Volume 8 2013

Published by :
IJDRI & Shannon Research
Adelaide, South Australia
ISSN : 1443-1475

IEJ welcomes practical and research manuscripts that focus on educational issues, provide a clear purpose and depth of discussion, and are presented in a straightforward style accessible to an international audience.

The issue of Bias
Avoid stereotyping on the basis of gender, race, or age. Accordingly, choose gender-neutral terms, such as sports person rather than sportsman describe the person, not the disability: for example a person with visual impairment rather than a visually impaired person use people of both sexes and vary the ethnicity of names avoid using the third-person singular pronouns he, his, and him by rewording the sentence with alternatives such as they or their, he or she, his or her, or him or her.

Copyright and Permission

CONTENT

International Education Journal, 2013, Volume 8, 110-126.
ISSN 1443-1475 © 2013.

Some problems in the analysis of cross-national survey data

John P. Keeves
School of Education, Flinders University *john.keeves@flinders.edu.au*

Petra Lietz
International University Bremen, Germany *p.lietz@iu-bremen.de*

Kelvin Gregory
School of Education, Flinders University *kelvin.gregory@flinders.edu.au*

I Gusti Ngurah Darmawan
School of Education, Flinders University *ngurah.darmawan@flinders.edu.au*

In this lead article three emergent problems in the analysis of cross-national survey data are raised in a context of 40 years of research and development in a field where persistent problems have arisen and where scholars across the world have sought solutions. Anomalous results have been found from secondary data analyses that would appear to stem from the procedures that have been employed during the past 15 years for the estimation of educational achievement. These estimation procedures are briefly explained and their relationships to the observed anomalies are discussed. The article concludes with a challenge to the use of Bayesian estimation procedure, while possibly appropriate for the estimation of population parameters would appear to be inadequate for modelling scores that are used in secondary data analyses. Consequently, an alternative approach should be sought to provide data on the performance of individual students, if a clearer and more coherent understanding of educational processes is to be achieved through cross-national survey research.

Cross-national research, survey research, secondary data analysis,
Bayesian estimation procedures, educational achievement

INTRODUCTION

As the number of school-aged children has grown rapidly world-wide and the demand for the provision of both primary and secondary education has increased at an even greater rate, it has gradually become essential to monitor educational standards. A little over 40 years ago the International Association for the Evaluation of Educational Achievement (IEA) was established and it set a pattern for the undertaking of the monitoring of educational achievement. Subsequently new bodies have been formed including the Programme for International Student Assessment (PISA) by the Organisation for Economic Cooperation and Development (OECD), the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) by the Ministers of 15 Sub-Saharan African countries, and there have been many independent studies conducted in single countries supported by the World Bank and other agencies. These different bodies have had similar objectives, but have gone about their work in different ways that have also changed over time. It would seem that these bodies have had four essential tasks to fulfil, although some studies might not have sought to undertake particular tasks.

1. The countries involved were to be ranked on educational performance of a particular kind, with appropriate estimates of standard errors.
2. Trends over time in educational performance of a particular kind were to be monitored and, where possible, factors influencing stability and change in performance were to be identified for each country.
3. Similarities and differences in both the factors and the patterns of factors influencing educational performance both within and between countries were to be identified, as well as the stability and change in the effects of these factors over time.
4. Research workers in each of the countries involved were to be trained in the conduct of studies concerned with assessment and the evaluation of educational achievement in order to plan for the raising of the standards of performance in each of the countries participating in the testing programs.

Each of the different bodies involved in the conduct of such testing programs have carried out these tasks to different extents in accordance with the financial resources available and the capacities of the research workers engaged in the programs to undertake the necessary analysis and training. However, the expansion of the combined efforts of the several bodies now involved has gone well beyond the initial expectations of the founders of IEA. As a consequence there has evolved gradually an understanding of the factors influencing both the provision of educational services and educational performance at the levels of students, classrooms and teachers, schools and school systems. Nevertheless, there is much more to be learnt and much more to be done in order to raise the standards of education in the schools of every country involved.

PERSISTENT PROBLEMS IN THE PRIMARY ANALYSIS OF DATA

Since the 1980s research workers have become very aware of certain persistent problems in the primary analysis of cross-national survey data. Several major problems have been encountered.

1. There was a partial failure of the models employed in the scaling and combining of test and questionnaire items to fit the data in particular countries, and under some circumstances there was a partial failure of tests of fit to detect a lack of fit because of circularity in some of the procedures used.
2. There was a need to conduct multilevel analyses at two or more levels (namely, students, classes, schools and strata or sub-systems) in order to model effectively the data recorded for both dependent and independent variables. There was also the need to calculate the appropriate errors of sampling in order to estimate accurately the statistical significance of the estimates of the parameters associated with such variables and the relationships between them (see Darmawan and Keeves, this issue, pp. 161-174, and pp. 175-190).
3. There were marked differences between countries in the best models that explained adequately the variability in the data associated with the variables under consideration. While there are sometimes strong similarities between groups of countries, there are commonly marked differences both within and between countries in the effects of certain independent variables on certain key dependent variables that lead to imposing serious limitations on the generalisations that can be drawn from the analyses (see Gregory, this issue, pp. 151-160) and (see Skuza, this issue, pp. 191-208).
4. Difficulties were encountered in the use of rotated test and questionnaire items when attempts were made to extend the coverage of different aspects of the school curriculum through

increasing the numbers of test and questionnaire items administered, without imposing too great a burden on individual students.

5. There were problems in the identification of appropriate models for combining data obtained from the administration of test and questionnaire items to form meaningful and consistent composite measures for the latent variables under consideration.

Since the 1960s there has been an ongoing debate about these issues in the universities and institutes engaged in educational research related to these assessment programs that have led to marked advances in the analysis of data in the field of education. These advances have gradually spread more widely to such fields as forestry, genetics, public health and the social and behavioural sciences. These developments include the techniques involved in structural equation modelling (eg. LISREL, PLSPATH, STREAMS, MPlus), multilevel analysis (e.g. HLM, MLwiN, MPlus) and measurement (e.g. Quest, RUMM, Bigsteps, ConQuest). New procedures to reduce the errors of measurement in population estimates have also emerged from the Educational Testing Service and Boston College in the United States and the Australian Council for Educational Research that have involved the use of conditioning and plausible values, which have remained obscured from a wider less technical audience until more general papers have been written recently by Adams (2005) and Wu (2005) from the Australian Council for Educational Research, and the University of Melbourne. These papers have made more readily accessible certain ideas associated with the procedures being widely employed in cross-national testing programs at a time when there is some concern about certain anomalous results that are encountered in the secondary data analysis of cross-national data.

SOME EMERGENT PROBLEMS IN THE SECONDARY ANALYSIS OF DATA

Ongoing efforts have been made over time to improve the quality of assessment and evaluation procedures. They have included: (a) the development of instruments that go beyond the administration of multiple choice test items to employ constructed response items with partial credit being given for less than complete responses; (b) the raising of the level of response rates both within and across schools; (c) the making of effective provision for the estimation of missing data, so that the designed samples may be adequately filled; (d) greatly improved methods of statistical analysis to estimate both direct and indirect effects of variables that influence educational outcomes at the between student, between classroom, between school, and between system levels; and (e) the use of meta-analytic and trend analysis procedures (see, Chiu and Khoo, 2005) to combine results from different countries, different studies and over time in order to develop a better understanding of stability and change in educational provision around the world.

Nevertheless, three highly anomalous findings have emerged from the secondary analysis of data that cast serious doubts on the strength and appropriateness of certain procedures that are currently being widely employed: (a) to provide for different tests being administered to different students,
(b) to compensate for missing data, and (c) to remove or reduce measurement error in order to improve the accuracy of population estimates. These procedures have been developed to overcome the limitations of test and sample design and response measurement. These three anomalous findings are considered briefly and in turn.

1. Meta-analysis of gender differences in reading achievement

In order to examine the gender differences in reading achievement at the middle secondary school level across a wide range of countries, Lietz (this issue, pp. 127-150) carried out a meta-analysis study that involved 147 data sets from a large number of testing programs including the IEA Reading Comprehension Study in 1970/71, the IEA Reading Literacy Study in 1990/91, the National Assessment for the Evaluation of Educational Progress Studies (NAEP) in the United

States from 1971 to 2003, the Programme for International Student Assessment (PISA) in 2000 in 43 countries, as well as the Australian ASSP and LSAY studies and many other smaller investigations. The meta-analysis was carried out using hierarchical linear modelling (HLM) procedures.

In the analysis, the outcomes examined were effect sizes, with their estimated errors, using a procedure advanced by Raudenbush and Bryk (2002, p. 209). The striking and anomalous finding was that the estimated effect size was substantially higher for the PISA studies (ê = 0.24) with similarly high values for the NAEP studies for the most recent decade (1992-2003), but not before that period. In contrast, studies prior to 1992 showed considerably lower effect sizes as reflected, for example, in the estimated effect size for the Reading Literacy Study, conducted by IEA in 1990-91 that was not significant (ê = 0.02). In general, in these more recent studies the girls were outperforming the boys with estimated effects that were noticeably greater than would be expected by chance (see Lietz, this issue, pp. 127-150). It is possible that these findings reflected the influence of cultural change, not only in the United States and Australia, but also in the 40 and more other countries of the world that have participated in the PISA and IEA studies. However, it is also possible that these effects arise from the item selection procedures employed to avoid gender bias, or from the procedures used for scaling and compensating for missing data and improving the accuracy of the national estimates of performance.

2. Mathematics Proficiency of Secondary School Students in South Africa

Howie (2002) undertook a secondary analysis of data on mathematics proficiency conducted as part of the Third (Trends in) International Mathematics and Science Study-Repeat (TIMSS-R) in South Africa in 1998/1999 under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). The striking finding was that the highest achieving students, coming largely from the Western Cape Province, scored approximately 100 score points or one student standard deviation below the international average of 487 score points. The Western Cape Province was the wealthiest and most urbanised province in South Africa, where English was widely used. While only about seven per cent of the total sample spoke English as the main language in the home, the students from Western Cape Province were in the main English speakers at home. The remarkably low level of proficiency of these students in Western Cape Province, indicated that they were probably about three years[1] behind the international average in their level of achievement in mathematics. This finding suggested that a highly anomalous result existed that demanded rather more thorough investigation, not merely of cultural effects, but also into how these scores were estimated.

3. Trends in Bulgarian Eighth Grade Mathematics Performance from 1995 to 1999

Gregory and Bankov (2005) undertook a secondary analysis of the performance in mathematics achievement of eighth grade students in Bulgaria between 1995 and 1999 in the Third (Trends in) International Mathematics and Science Studies (TIMSS and TIMSS-R). In 1995, the Bulgarian students had a mean achievement of 527 that was above the international mean of 500. However, in 1999 the level of Bulgarian mathematics achievement fell to 511, which was still significantly above the international average. This involved a decline in performance of about one-sixth of a student standard deviation or approximately half a year of schooling that not only was statistically significant but was also of considerable practical significance. The possibility of unknown problems in the sample design could not be ignored. Nevertheless, a decline in performance of this magnitude could well indicate some anomaly in the scaling procedures used, or a marked change in the structure of the tests employed that was associated with an incompatibility of the

[1] This estimate for Mathematics achievement is obtained from Afrassa and Keeves (1999).

items in the different test booklets with the Bulgarian mathematics curriculum, where much of the content tested had been taught to the Bulgarian students between one to four years before the eighth grade. However, it was also possible that changes made in 1997 to the Bulgarian school system contributed to this anomalous result, and consequently the findings of the study raised politically sensitive issues. Nevertheless, in view of the other two anomalies discussed above, this result might indicate that a problem existed in the procedures used in the scaling of achievement data that warranted further examination.

SOME GENERAL ISSUES IN THE ANALYSIS OF DATA

There are several issues in the analysis of data that emerged over the period during which cross-national studies of educational achievement had been carried out, that need to be recognised and understood before the three anomalous effects considered above can be discussed and possibly addressed.

1. The effects of bias due to missing data

The occurrence of substantial missing data at the student and school levels, in general, would have the effect of introducing bias to both the estimates of the mean level of achievement and the variance. Such bias would be likely to inflate the mean level of achievement and reduce the variance, because it would seem likely that some lower performing students and schools from the designed samples would fail to participate in the study. These sources of bias would serve not only to distort the estimated level of performance, but also to reduce the capacity of the analysis of variance procedures employed in subsequent analyses to detect effects.

2. The effects of non-normal multilevel generating distributions of data

For the achievement test outcome variables and the indicators of attitude and the contextual variables formed by the summation of scores or by principal component procedures the underlying generating distributions would be likely to involve approximately normal distributions. However, there would be many key variables that would have to be included in the analyses of the data, such as the sex of a student and school type that could not be considered to be normally distributed. The failure to have underlying normal distributions would not only be likely to influence the use of significance tests, but could also influence the use of certain maximum likelihood estimation procedures. Sometimes, however, appropriate transformations could be used. Nevertheless, the underlying normality of the generating distributions would require very careful consideration, if and when maximum likelihood estimation procedures were employed.

3. The level of analysis problem

Only in the period since 1985 have effective analytical procedures been made available for an effective consideration of the multilevel analysis problem that has existed in educational research studies, where data were collected from students nested within schools. While over the past 40 years procedures have been employed to make some allowance for this aspect of the study and sample design in significance testing, it has not been possible to provide for the clustered sample design in the estimation of effects at appropriate levels until very recently. Even within the more highly developed countries there would sometimes, but not always, be substantial problems arising from the design of the sample, where these effects differed markedly between variables. Moreover, in many developing countries that currently participate in the IEA and PISA studies there would be very substantial design effects not only associated with individual schools but also associated with clearly identifiable regions and types of schools, as for example, academic, comprehensive and technical schools. These would require the use of a third level of analysis for

the appropriate estimation of statistical significance and the unbiased estimation of effects, since these school effects would be fixed effects and systematic in nature, and not random effects.

4. Bivariate and multivariate analysis

In a major debate that occurred 40 years ago the analysis of data in cross-national achievement studies shifted from an examination of bivariate relationships using simple analysis of variance procedures to an examination of multivariate relationships using regression procedures. Subsequently, a further development in the use of regression procedures led to the estimation of not only the direct effects of variables, but also the indirect effects of variables on the outcomes under consideration. However, the simplicity of bivariate relationships would seem still to have its appeal, whereas the real world of schooling would appear to be built out of a complex network of direct and indirect effects that required careful modelling at different levels of analysis. All efforts involved in the development of carefully constructed and trialled questionnaires would ultimately be wasted, if only bivariate relationships were examined and if multilevel and multivariate path models were not constructed to represent and tease out the effects of factors that influenced the educational outcomes within a particular education system and between educational systems (see, Chiu and Khoo, 2005).

5. The specification of regression models

The development and testing of regression models clearly would demand a thorough and systematic multilevel and multivariate analysis of variance using regression or maximum likelihood estimation procedures, with full recognition that each education system was likely to be very different from its neighbouring systems, because of the historical and cultural factors that had led to the formation in each country of a unique education system. While the questionnaires employed in the cross-national achievement surveys have sought to obtain meaningful data from students, teachers, and school principals, the questionnaires have frequently been returned with substantial missing and inconsistent information. Consequently, appropriate regression based procedures would be required to provide estimated values of missing data in those questionnaires where such data were missing or were inconsistent.

With the increasing number of countries involved in the surveys it is becoming more and more difficult to develop questionnaires that obtain meaningful data from the wide range of countries involved. Moreover, while in some highly developed countries there is little variability between schools in both their characteristics and the levels of achievement of their students, for many developing countries there are frequently wide disparities both in characteristics and levels of achievement. As a consequence, there are major differences between countries in the structures of the models that are constructed to explain optimally the differences between schools and students in their levels of achievement. Furthermore, there are likely to be large differences between countries in the explanatory power, in terms of proportion of variance explained, in the optimal models developed to account for variation in achievement outcomes.

6. The construction of tests and the sampling of test items

A major problem in the conduct of a testing program is that there is a relatively small limit to the number of test and questionnaire items to which a student can be asked to respond. This demands that in any content domain each student is required to answer only a sample of the test items that are employed to cover the content domain with adequate content and construct validities. A balanced incomplete block (BIB) design is currently widely used and compensation is made in estimating test scores not only for missing data but also for the different tests answered by different students.

Weiss and Yoes (1992) stated that there were two major approaches for estimating student performance, namely, the maximum likelihood method and the Bayesian estimation method. The maximum likelihood approach gave rise to two commonly used estimation procedures, either the Rasch (one parameter) modelling of the data or the three-parameter modelling of the item data collected in the testing program. The former modelling procedure provided measures that were said to be independent of the items sampled and the persons involved in the calibration of the scale of measurement. The former procedure also demanded that both the items and the persons tested must satisfy strict requirements of uni-dimensionality. The latter procedure, while claiming to be more accurate, has generally been found to be less robust.

In order to improve estimation and to compensate both for missing data and the BIB spiralling of the tests, an additional step beyond the maximum likelihood method involving the Bayesian estimation procedure has since 1992 been widely employed. In order to improve further the estimation, instead of relying solely on one estimated value, five plausible estimates have commonly been generated for subsequent analysis. These plausible values have been provided through the use of a so-called 'conditioning' procedure not only to replace the missing data, but also to replace all achievement test data, in order to improve both the effects of BIB spiralling, as well as to reduce the errors of measurement. It is argued in this article, that from the employment of the Bayesian estimation procedure that involves the formation of a prior distribution of estimated performance, the anomalous findings considered above may well arise.

A DISCUSSION OF PROCEDURES FOR ESTIMATING STUDENT PERFORMANCE

Adams (2005) and Wu (2005) have, with considerable clarity, in their published articles addressed these problems, in ways that were complementary and very informative. It was clear that because of BIB spiralling simple procedures that involved raw scores could no longer be employed to provide precise national estimates of the mean level of performance. However, maximum likelihood estimation procedures, involving either Rasch measurement or the three-parameter model could be used. Several other scoring procedures could also be used that were generally grouped within the two categories that Weiss and Yoes (1992) specified. These are listed below.

Maximum Likelihood Estimates (MLE)

One of three alternative procedures could be used:

(a) the Rasch model using the ConQuest, Quest, Bigsteps or RUMM programs that employed the one parameter measurement model;

(b) the three parameter model using BILOG or SAS/ETS enhance programs that employed the three parameter model; or

(c) the Weighted Maximum Likelihood Estimates (WMLE) obtained using the Warm Likelihood Estimation (WLE) procedure in which the maximum likelihood estimates for each individual were weighted by the information function for the set of items to which each individual had responded (Warm, 1989).

Bayesian Estimates

Two procedures could be used:

(d) the Plausible Values (PV) procedure that involved the use of five values which were sampled from a posterior distribution of the score for each individual; or

(e) the Expected A-Posterior Estimate (EAP) that involved calculating the mean of the posterior distribution for each student.

Three important statements were made about the different estimation procedures by Adams (2005) and Wu (2005) that were associated with the use of the Rasch model.

1. *Bias involved in estimates*

   All estimation procedures provided unbiased estimates of the mean score for the group.

2. *The maximum likelihood estimates*

   The maximum likelihood estimation procedure provided an unreduced estimate of the variance of the scores of the group. This was simply because no provision had been made to reduce the variance of the group scores that arose from errors of measurement.

3. *The Warm likelihood estimates*

   The Warm (1989) or weighted likelihood estimation procedure reduced the variance of the scores of the group by weighting each individual maximum likelihood estimated score distribution by the information function for each point estimate on the score distribution. The information function was defined by Fisher (1922) as the reciprocal of the precision with which a parameter was estimated. This information function was related to the square of the slope of the item characteristic curve at different points on the curve, and was standardised by dividing by the conditional variance:

$$I(\theta, u) = \frac{dp}{d\theta}^2 \Big/ \text{conditional variance,}$$

   where $\frac{dp}{d\theta}$ = the slope at different points on a test characteristic curve, and $I(\theta, u)$ = the information function.

   Combining the likelihood distribution function (MLE) with the information function to form their product yielded the Warm likelihood function (WLF).

   The maximum value of the Warm likelihood function is referred to as the 'Warm likelihood estimate (WLE)' or the 'weighted likelihood estimate'.

Wu's simulation study

Wu (2005) has reported the results of a simulation study that provided information on the characteristics of the different estimates for both 3-item tests and 20-item tests where the generating distribution was N(0, 1) for the 3-item tests and N(2, 1) for the 20-item tests. These results are given in Table 1.

Wu (2005) drew the following conclusions from her simulation study.

1. MLE values might show some bias in mean values and greatly over estimated the variance of the generating distribution that was not adequately adjusted by a reliability correction. Thus variance associated with measurement error was clearly present in MLE values.
2. WLE values showed little bias in the mean values and overestimated the variance of the generating distribution. However, the Warm estimating procedure removed some but not all of the variance associated with measurement error. The reliability correction reduced the variance well below the expected value.
3. The plausible values (PV_1 to PV_5) were constructed to have an unbiased mean and an appropriate variance. It was argued that the measurement error had been removed by the conditioning process.

4. The estimated posterior (EAP) values that were formed as the mean of the plausible values were unbiased, but had as might be expected, substantially reduced variance. This variance was well adjusted by the reliability correction.

Table 1. Comparison of estimates for simulated 3-item test and 20-item test.[a]

	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	GV_c
3-item Test Estimated Mean	-0.002	0.002	-0.002	-0.000	-0.004	-0.003	-0.002	-0.003	0
Standard Error	(0.030)	(0.039)	(0.036)	(0.041)	(0.042)	(0.042)	(0.041)	(0.041)	
Estimated Population Variance	1.950	2.350	0.359	0.995	1.004	1.002	1.004	1.001	1
Standard error	(0.263)	(0.178)	(.061)	(0.113)	(0.108)	(0.112)	(0.113)	(0.109)	
Corrected Value [b]			0.99						
	WLE	MLE	EAP	PV1	PV2	PV3	PV4	PV5	GV_c
20-item Test Estimated Mean	1.966	2.117	2.002	2.002	2.002	2.000	2.003	2.003	2
Standard Error	(0.031)	(0.031)	(0.032)	(0.035)	(0.033)	(0.033)	(0.032)	(0.135)	
Estimated Population Variance	1.332	1.657	0.683	1.003	1.005	1.007	1.003	1.003	1
Standard Error	(0.051)	(0.056)	(0.047)	(0.059)	(0.061)	(0.061)	(0.063)	(0.059	
Corrected Value [b]	0.72	0.89	1.01						

[a] adapted from Wu (2005); [b] correction made for unreliability of estimates; [c] GV – Generated Values

Bayesian estimation, conditioning, and plausible values

The Bayesian estimation procedure that involved the construction of the prior distribution and its use in modifying the likelihood score distribution to form the posterior distribution has been referred to as a 'conditioning' procedure. Conditioning not only provided estimates for any missing scores, but it also refined the maximum likelihood estimates for all individuals. In addition these estimates were also replaced by five plausible values as well as an EAP estimate that was the mean of the five plausible values and the mean of the posterior distribution.

Adams (2005) and Wu (2005) presented evidence to support the case both for the use of plausible values and conditioning that would appear to have a high degree of credibility. Nevertheless, it is contended that through their cursory treatment of the construction of the prior distribution they failed to emphasise a potential shortcoming associated with the use of Bayesian estimates. Wu (2005, p. 125) recognised that a degree of bias might be associated with the estimates of population regression coefficients in the following words.

> The degree of bias of the regression coefficients will depend on test length and the partial correlation between the variable of interest and the latent variable, after controlling for any conditioning variables that were used. When a regression analysis is run using plausible values generated with a model that did not include the regressors, it is said that *model unspecification* has occurred. (Wu, 2005, p.125)

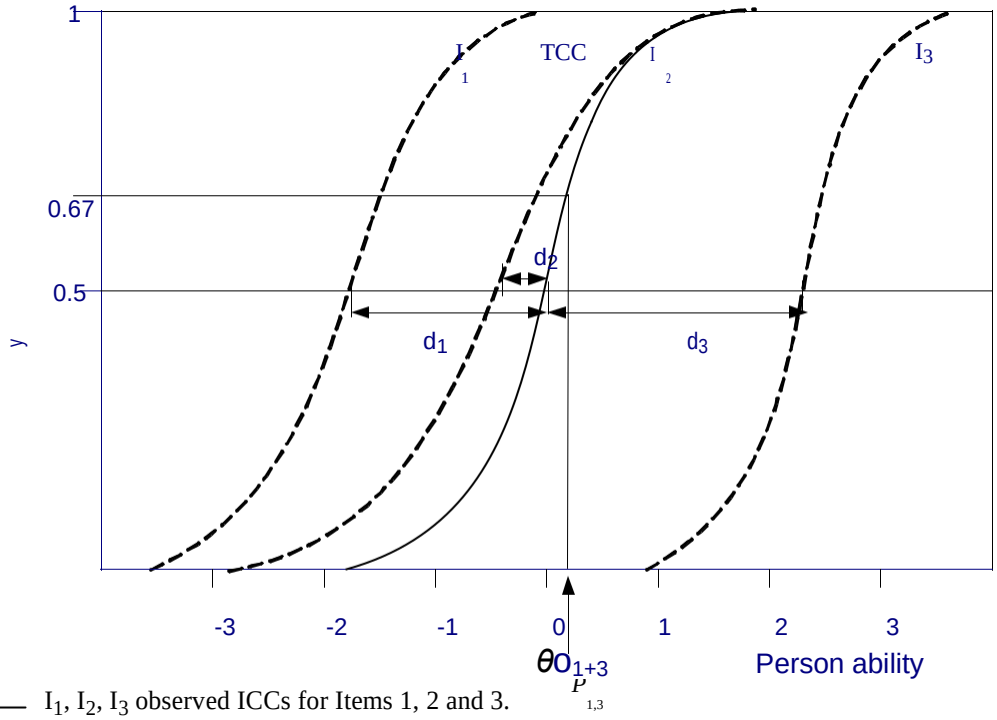
These qualifications are important but are clearly not enough and can be said to be both incomplete and inadequate. It is necessary to support the authors' assertions by a discussion of the

three anomalous cases that have been observed in secondary data analyses. It is also necessary to ask, with an understanding that has been developed from reading the papers by Adams (2005) and Wu (2005) and from experience in the primary and secondary analyses of cross-national data, whether suggestions can be advanced as to how the three observed anomalies might have arisen from the use of WLE, EAP or PV values. It is recognised that difficulties are encountered in the data analyses in such studies, and that the attempts made to provide for missing data and measurement error are necessary and desirable. Moreover, the authors apologise for failing to test fully their ideas by undertaking further analyses. However, before considering these anomalies, it is necessary to explain in greater detail the estimation procedures that are being employed in these studies.

A diagrammatic treatment of the estimation procedures

In the section that follows a diagrammatic explanation is presented of the estimation procedures without discussing these procedures using mathematical symbols. The figures are presented as illustrations of certain effects and are not derived from simulation or the use of particular measurements.

In Figure 1, three item characteristic curves are shown for Items 1, 2 and 3, the combined test characteristic curve for Items 1, 2 and 3 that were attempted by Person $P_{1,3}$, and the maximum likelihood estimate (MLE) for p the probability of a correct response for Person $P_{1,3}$, who responded correctly to Items 1 and 3, but not to Item 2.



I_1, I_2, I_3 observed ICCs for Items 1, 2 and 3.

TCC the observed test characteristic curve for Item 1, Item 2 and Item 3.

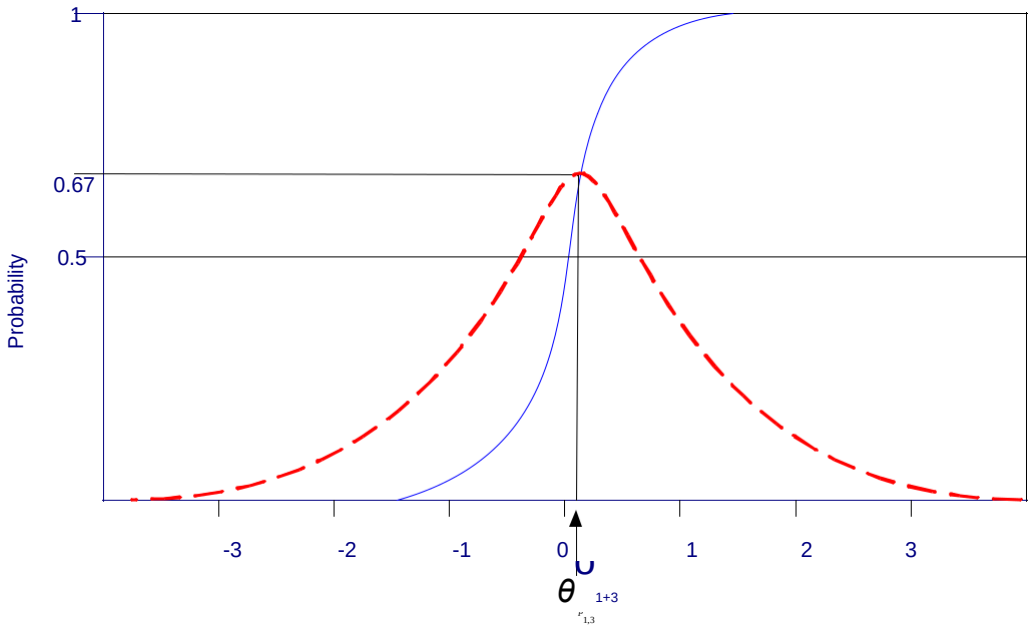
Where d_1, d_2 and d_3 are the item difficulty values and $d_1 + d_2 + d_3 = 0$.

Figure 1: Item characteristic curves for Items 1, 2, and 3 with I_1 and I_3 answered correctly

Since Person $P_{1,3}$ answered items I_1 and I_3 correctly with $p = 0.67$ a score of $\theta_{P_{1,3}}$ can be estimated. The zero for person ability is set at the average difficulty level of the three items when p = 0.5.

In Figure 2 it is shown how the response or likelihood distribution curve is trimmed to remove, in part, measurement error by weighting the likelihood function by the information function to provide a more precise estimate of the score of Person $P_{1,3}$, with reduced variance.

The likelihood distribution function is shown for Person $P_{1,3}$ who responded correctly to Item 1 and Item 3. The person's response or likelihood distribution function is combined with the information function to form the Warm likelihood function.



observed test characteristic curve for Items I_1, I_2 and I_3
information function (IF) for Items I_1, I_2 and I_3
response distribution function (MLE) for Person $P_{1,3}$
Warm likelihood function (WLE) for Person $P_{1,3}$ with ability score $\theta_{P1,3}$

Figure 2: Response function curves for Items 1 and 3 answered correctly

It can be seen that the θ values for MLE and WLE remain close together showing little bias. However, the Warm likelihood function has less variance than the response distribution function for Person $P_{1,3}$, because it is formed by combining the information function with the response distribution function.

If missing data need to be imputed a prior distribution is clearly required, and consideration must be given, as to how best to produce this prior distribution.

A commonly used procedure is to employ a normal distribution N(0,1) as the prior distribution for each individual person and to impute the missing test score. However, it is also possible to construct a regression equation that best predicts the observed score for that individual using all known information about the group to which that individual belongs and to use this regression

equation as a prior distribution in a normalised or standardised[2] form in order to predict the score for the individual for whom a test score is missing. When some information is known about the individual the use of the prior distribution in this way to predict the missing score involves the use of Bayesian estimation procedures. If no information is known about the individual, scores are obtained at random using the posterior distribution for the group to which the individual belongs.

An extension of this principle is said to 'improve' or 'condition' the data by estimating the scores of all persons irrespective of whether or not their test scores are missing, and whether or not any other data are missing. In this estimation process the selection of a single best estimate proves inadequate, and the procedure currently adopted is to choose five estimates at random from the posterior distribution for the individual that is a combination of the likelihood or response distribution for the sub-group to which the individual belongs, and the normalised prior distribution, obtained by regression analysis procedures. If no specific information is known about the individual, the prior distribution represents the group and is based on the characteristics of the group to which the individual is said to belong, and scores can be estimated from the posterior distribution for the group. Clearly, at least five estimated scores are better than one. Moreover, because the posterior distribution is conditioned by the prior distribution to reduce measurement error and if all estimates are selected randomly from the posterior distribution, then the scores obtained follow the posterior distribution. Since the posterior distribution is a combination of two distributions, the scores that arise from the conditioning procedure form a distribution with reduced variance.

This procedure is presented in diagrammatic form in Figure 3.

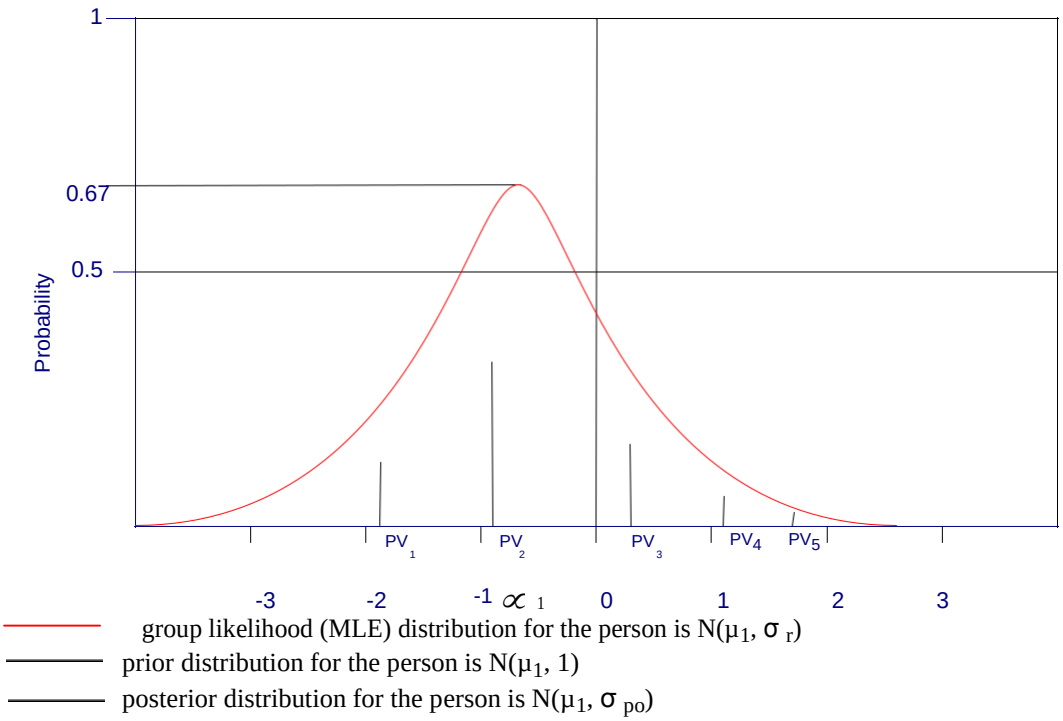


Figure 3: Bayesian Estimation in an ideal case for a specific person.

[2] A normalised or standardised distribution has a known mean of zero and a standard deviation of 1, and has a distribution that tends towards normality under the central limit theorem.

Consider the case of a person for whom a score is not available, but some information is known about that person and about the sub-group to which the individual person belongs.

The process of Bayesian estimation is shown for this specific person in which the sub-group likelihood response distribution is a $N(\mu_1, \sigma_r)$ distribution, and the prior distribution after regression analysis for that person is given by an approximately normal $N(\mu_1, 1)$ score distribution. The likelihood response function is then weighted or multiplied at each level of θ by the corresponding value of the prior distribution to obtain a new likelihood function referred to as the posterior distribution. This posterior distribution for the individual persons is a $N(\mu_1, \sigma_{po})$ distribution. Five plausible values are then chosen at random from this posterior distribution for the missing data where some information about the individual person is available, and are shown as PV_1 to PV_5. Thus where information is known about the individual, that information is used to obtain the five plausible values. The expected posterior estimate (EAP) is the mean of the five plausible values that are obtained for each individual. Where no information is known about the individual, the prior distribution for the sub-group to which the individual belongs is used.

In Figure 4 the process is displayed of Bayesian estimation in the case of a low performing group of students who fit the regression model developed less than adequately and the group distribution exhibits positive skew. The prior distribution for the group is estimated from regression analyses and also exhibits positive skew. Consequently, the mean value of the posterior distribution for the group is likely to be seriously biased.

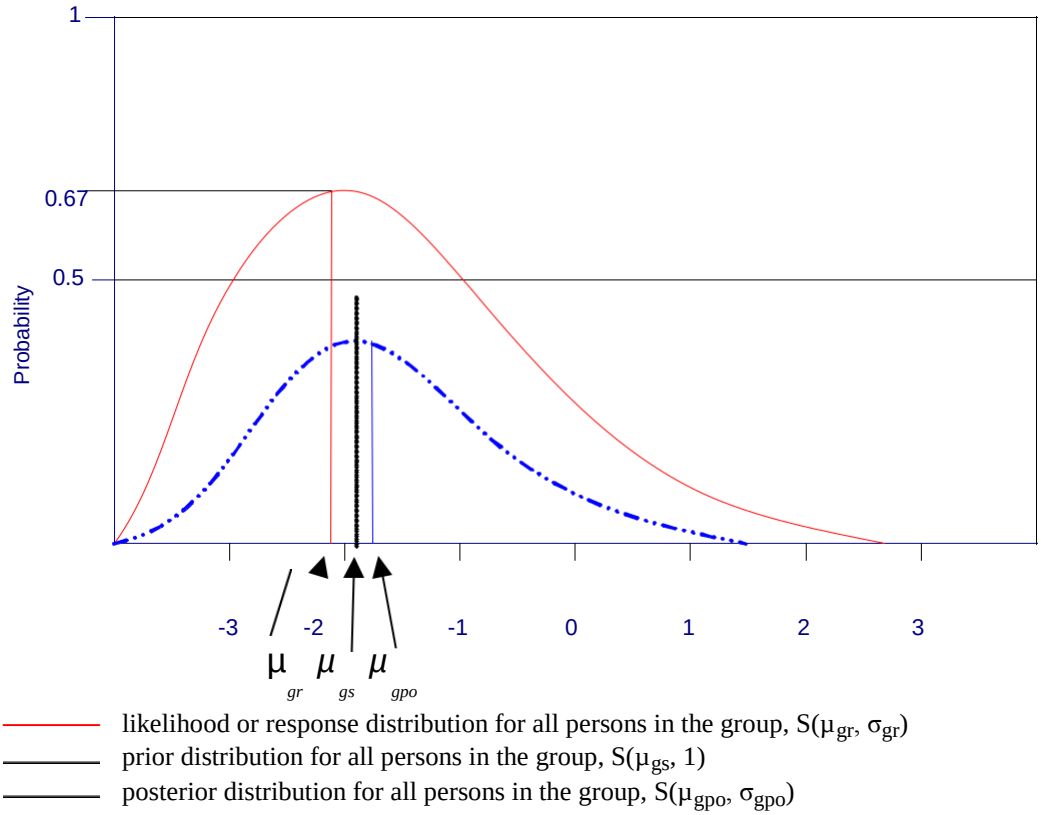


Figure 4: Bayesian estimation for a low performing group showing the response distribution for the group and the prior and posterior distributions

The major problem with the conditioning process and the use of the Bayesian estimation procedure is that if the prior distribution is not estimated well poor plausible values are obtained

for the individuals in the sample as well as for subgroups of individuals although the population estimates may be adequate.

OUR THOUGHTS ON THE ANOMALIES THAT WE HAVE OBSERVED

Meta analysis of gender differences and reading achievement

In the conduct of meta-analysis there was interest in the changes in important effects over time and in the differences that arose across different cultures and different education systems. It would seem likely that the gender effects that Lietz (this issue, pp. 127-150) compiled were derived before 1992 from raw score and Rasch scaled population estimates for male and female students. Since 1992, Bayesian procedures have been used in the estimation of both means and variances. These estimates would have reduced variance, as measurement error would have been removed. This reduction in variance would have given rise to larger effect sizes. Under these circumstances any attempt to undertake the meta-analysis of estimated effects might need to distinguish between, before the use of Bayesian procedures and after the use of Bayesian procedures. Other procedures such as raw scores and likelihood estimation procedures should yield similar mean values for male and female groups, where large groups were involved, but with much greater variances and consequently greatly reduced effect sizes.

Mathematics Proficiency of Students in South Africa

In the analysis of the data for the South African sample of students who were tested for proficiency in mathematics, there would be little doubt that any regression analyses of the data carried out would show that the characteristics of the South African sample were very different from other samples involved in the TIMSS studies. As a consequence the prior distributions used in the Bayesian estimation of the South African scores would differ markedly in variance from other national samples, probably casting serious doubts on the use of these procedures in the analyses of these data. What probably happened in the conditioning of the South African data is displayed in Figure 4 where the small group of higher performing English speaking students would be pulled back markedly towards the lower end of the scale in the Bayesian estimation process and conditioning operation.

Trends in Bulgarian Eighth Grade Mathematics Performance.

The possible explanation of the significant decline in mathematics achievement over the short time-span of four years probably lay in a shortcoming in the construction of the regression model that was used as the prior distribution in the Bayesian estimation of the posterior distribution from which the national mean value was estimated. It would seem possible that a variable that involved the changed structure of the strata employed for the sub-systems into which the schools were grouped was not examined in 1995 and 1999 in appropriate ways in the regression analysis to form the prior distribution for the Bayesian estimation procedure. However, it would also be possible that the different test booklets that were employed on the two occasions differed in important ways with respect to the mathematics curricula of Bulgarian schools at the eighth and lower grades. A consequence of this would be that the original score distributions were influenced differently on the two occasions by the lack of match between the curriculum and the different test booklets that were used to sample and estimate student achievement in mathematics.

Further comments

In Bayesian estimation the likelihood or response distribution is modified by the prior distribution to yield the posterior distribution of scores to differing extents for different countries, different sub-systems and different individuals. If the prior distribution reflected adequately the original

likelihood or response distribution it would seem that little distortion would be likely to occur. However, if the prior distribution did not reflect adequately the likelihood or response score distribution, as a consequence of important factors not being included in the prior distributional model, then the effects of those factors would most likely tend to disappear from the scores that were made available for secondary analysis.

Wu (2005), as stated above, warned against model mis-specification when regression analyses were undertaken to form the prior distribution. This suggested that any subsequent construction of explanatory models in secondary data analysis would largely be a waste of time and effort because of specification problems in the prior distribution and the magnitude of such effects would remain unknown.

Warm likelihood estimates (WLE) have also been provided in the data files to enable secondary data analyses to be undertaken with scores that were not modified by the effects of the prior regression-based distribution. However, little appears to be known about the effects of trimming the variance of the scores by the procedure proposed by Warm (1989). The use of this procedure must be expected to have consequences for the estimation of the effect sizes.

Further possibilities associated with the Bulgarian analysis could have arisen in two different ways. The regression analyses that were carried out in the forming of the prior distribution were most likely undertaken only at the student level through the use of the general linear model. If a two level model were used it might be possible to provide for effects at the student and school levels. However, many national school systems had strikingly large differences between regions or provinces and states, as well as school types, and the use of at least a three level model would seem to be required. In the formation of the posterior distribution it should be recognised that 'what you get out is strongly related to what you put in'. Unfortunately, little information has been made available on the nature of the variables employed in constructing the prior distribution in different countries and for different groups of students or on the amounts of variance involved at the different levels of the data. Furthermore, there has been little information provided on the nature and extent of differences between the different national education systems with respect to the strongest factors that were associated with the development and construction of the prior distribution within each system that had such a pivotal role in the conditioning process.

The BIB spiralling procedures that are built on the use of eight different test booklets and that serve to increase the range of content which can be assessed, employ items that are frequently clustered under a common stem. Thus the range of content assessed by each booklet is very limited. Our secondary analyses have shown that the different booklets operated very differently across countries in sampling student performance probably because of differences across countries in the structure of the curricula under survey. The relationships between the content of the items in the test booklets and the opportunity that the students in different countries had to learn that content and the performance of students in different countries has been a controversial issue over the past 40 years during which cross-national assessment programs have been operating. Unfortunately, little progress would appear to have been made over the years in the examination of curriculum design and time allocated to learning the content assessed by the test items and their effects on learning outcomes. These aspects are possibly involved in the anomalous effects recorded over time in the Bulgarian analyses of the TIMSS data.

CHALLENGING THE USE OF THE BAYESIAN PROCEDURE

Michell (1986, 2000) has raised questions about the nature of measurement in the behavioural and social sciences identifying the three theories of representational, operational and classical measurement. It would seem that, in practice, elements of all three theories are generally involved.

Moreover, measurement provides among its other functions, a structure for the use of mathematical symbols, ideas and relationships. Not only must the measures bear a representational relationship to qualities, but the measures must also express the relationships between such qualities that involve operational purposes.

Nevertheless, in the social and behavioural sciences abstract qualities are involved and measures of these abstract qualities are sought. Furthermore, information on how much of an abstract quality is involved, namely its measure is required. In addition, information about an abstract quality can only be obtained through the interaction of people with tasks that are associated with the quality under consideration. Consequently, in order to estimate the ability of a person all that can be observed is performance on a task. The difficulty of the task must both be sampled on multiple occasions and in multiple situations that involve probabilistic or stochastic relationships. Thus, several different sources of error must be taken into consideration. The discussion in this article is primarily concerned with those sources of error that are associated with performance errors that arise from:

(a) variability in the performance of the person involved,
(b) variability in the tasks being performed, and
(c) variability in the observation of performance on the tasks.

In general, tasks, observations and persons or cases are sampled, and thus sampling errors are also involved. The procedures adopted by Adams and Wu in their work seem to be directed towards certain operational aspects of measurement to the exclusion of other representational aspects, on the assumption that a so-called 'true' value is capable of being estimated. Such a 'true' value is unknowable in the social and behavioural sciences.

It is argued in this article that the work of Adams and Wu fails to satisfy the requirements of both representation and operation as they move beyond classical approaches. The use of plausible values is not appropriate for estimating the scores of individual students and certain subgroups of students. The plausible values and the EAP values are better suited for estimating the performance of a population. Consequently, it is also argued that other ways must be sought to allow for uncertainty and the use Bayesian estimation procedures should be rejected. Other error estimation procedures are available. For example, bootstrapping or jackknifing of items and persons with respect to their primary sampling units can be used to provide estimates of measurement error in the same way as bootstrapping and jackknifing are used to provide estimates of sampling error. This, however, seems to require a major rethinking of the strategy of data analysis that has evolved around the use of Bayesian estimation methods. These estimation procedures although apparently effective for the better estimation of population parameters, are made at the expense of individual and sub-group estimates, which are essential for the examination of multivariate and multilevel models. While information on trends in population mean values over time is of importance in the monitoring of educational outcomes, the development of a clearer and more coherent understanding of educational processes and how these change over time was not only the goal set by the founders of IEA, but remains today the most challenging task for those who believe in the importance of increasing the effectiveness of education and its contribution to human development.

REFERENCES

Adams, R. J. (2005) Reliability as a measurement design effect, *Studies in Educational Evaluation*, 31, (2/3), 162-172.

Affrassa, T.M. and Keeves, J.P. (1999) Changes in students' mathematics achievement in Australian lower secondary schools over time, *International Education Journal,* 1(1), 1-21.

Chiu, M.M. and Khoo, L. (2005) Effects of resources, distribution inequality, and privileged bias on achievement: Country, school, and student level analyses. *American Educational Research Journal*. 42(4), 575-604.

Fisher, R. A (1922) On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London* A, 222, 309-368.

Gregory, K.D. and Bankov, K. (2005) Exploring the Change in Bulgarian Eighth Grade Mathematics Performance from TIMSS 1995 to TIMSS 1999. In T. Plomp and S. Howie (Eds.), *Contexts of Learning Mathematics and Science Lessons Learned from TIMSS*. London: Routledge.

Howie, S. (2002) *English Language Proficiency and Contextual Factors Influencing Mathematics Achievement of Secondary School Pupils in South Africa.* The Hague: CIP-Gegerones Kononklyke Bibliotheck.

Michell, J. (1986) Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin* 100(3), 398-407.

Michell, J. (2000) Normal science, pathological science and psychometrics. *Theory and Psychology*. 10(5), 639-667.

Raudenbush, S.W. and Bryk, A.S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd edition). Thousand Oaks, CA: Sage Publications.

Warm, T.A. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54(3), 427-450.

Weiss, D.J. and Yoes, M.E. (1992) Item Response Theory. In R.K. Hambleton and J.N. Zaal (1992) *Advances in Educational and Psychological Testing*: *Theory and Applications*. (pp. 69-95), Dordrecht, The Netherlands: Kluwer.

Wu, M. (2005) The role of plausible values in large-scale surveys. *Studies in Educational Evaluation.* 31, (2/3), 114-128.

International Education Journal, 2013, Volume 8, 127-149.
ISSN 1443-1475 © 2013.

Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view

Petra Lietz
International University Bremen, Germany *p.lietz@iu-bremen.de*

Results of a previous meta-analysis of gender differences in reading achievement at the secondary school level (Lietz, in press) showed significant differences between major assessment programs. Thus, the gender gap in favour of girls was more pronounced for the assessment programs conducted by the National Assessment of Educational Programs in the United States (NAEP), for the more recent assessment programs in Australia and the Programme for the International Student Assessment (PISA) conducted by the OECD. In contrast, no such effect was found for earlier studies conducted by the International Association for the Evaluation of Educational Achievement (IEA), namely the International Reading Comprehension Study 1970-71 and the International Reading Literacy Study 1990-91.

Hence, this article seeks to investigate whether or not an effect exists that could be associated with the time period in which a study was conducted. In other words, the article examines whether or not the reasons for the greater gender differences in more recent assessment programs might be related to the scaling of reading scores before and after 1992.

Reading achievement; scaling of scores; meta-analysis;
hierarchical linear modelling; gender differences

INTRODUCTION

The research reported in this article extends across two major areas, one content-related area, namely gender differences in reading, and one method-related area, namely meta-analysis. Each of these areas is discussed briefly below.

Gender Differences in Reading Achievement

The view prevails that boys perform better than girls in mathematics (Aiken and West 1991, Johnston and Dunne 1996, Husen 1967, Keeves 1988, Tracy 1987) and the natural sciences whereas the reverse holds in reading, social studies and languages (Dedze 1995, Plisko 2003, Thorndike 1973, Wagemaker et al. 1996). A closer examination of the research on reading, however, reveals that the matter is not as clear-cut as it might appear and that results can be grouped into two main categories: one showing evidence of girl's superiority over boys in reading achievement, and one providing little or no evidence of gender differences. Thus it can be argued that the research provides some support for the existence of a gender gap in reading performance in favour of female students, while some studies and reviews dispute this finding. However, these studies provide inconclusive evidence with regard to the extent of gender differences in reading at the secondary school level. Hence, a more systematic approach to integrating research findings,

namely statistical meta-analysis (Glass, McGaw and Smith 1981, Hunter and Schmidt 2004) is suggested and discussed in greater detail below.

Meta-Analysis

For 40 years and more, reports of research findings concerned with the magnitude of the difference between two means have recorded the size of an effect in terms of a standardised difference. This standardised difference was first referred to as an 'effect size' by Cohen (1969). The effect size was calculated by dividing the difference between the means of the two independent groups, by the pooled standard deviation of the two groups. Moreover, Cohen showed how it was related to the point biserial correlation coefficient, not only by multiplying the correlation coefficient by 2, when two large groups were of approximately equal size, but also by using another multiplying factor for unequal sized groups.

Subsequently, the term 'meta-analysis' involving an analysis of effect sizes was introduced by Glass (1976, 1977) to denote a systematic integration of research findings on a specific topic and has been developed further as an analytical technique (Rosenthal 1984, Hedges and Olkin 1985). The need for a more systematic way of integrating prior research than narrative research reviews was introduced as a reaction to criticisms aimed at the social sciences by funding agencies and the public as to whether or not any progress was being made in terms of establishing some statements of knowledge from the seemingly abounding and contradictory evidence generated from many research projects in the social sciences (Light and Smith 1971).

As Hunter and Schmidt (2004, p. 16) emphasised: "In many areas of research, the need today is not for additional empirical data but for some means of making sense of the vast amounts of data that have been accumulated." Moreover, they point out that the narrative integration of research findings has serious shortcomings in that this strategy of integrating research results often leads to different conclusions if done by different people. Statistical meta-analysis, in contrast, as a quantitative way of integrating research findings should lead to the same conclusion, regardless of the person applying the procedure.

Thus, the challenge in the social sciences, in general, and in educational research in particular, is to integrate systematically and quantitatively findings from the large number of research studies that have been undertaken in order to contribute empirically verified facts to the cumulative body of knowledge.

None of the meta-analyses undertaken to date have focused specifically on gender differences in reading. In addition, advances in hierarchical linear modelling (HLM) have occurred that allow for the clustered nature of meta-analytic data to be taken into account more appropriately. Thus, Raudenbush and Bryk (2002) argued that the main purpose of a meta-analysis was to examine the extent to which effects reported in the results of primary studies were consistent and to disentangle what part of the variance in study results was due to sampling error and what component was due to actual treatment implementation. As a consequence, Raudenbush and Bryk (2002) proposed an empirical Bayes meta-analysis as a special application of the two-level hierarchical linear model. In this model, the outcome variable, namely the effect sizes from the different studies, was allowed to vary randomly at the first level while, at the second level, study characteristics were used to explain possible differences in the outcome variable. In other words, the Level-1 analyses were aimed at investigating the extent of the variability in effects sizes of primary studies, while at Level-2 possible sources of this variation might be examined. This extension to two levels was based on the use of ordinary regression models in research synthesis proposed originally by Hedges and Olkin (1983).

In summary, meta-analysis is a systematic way to synthesise findings of research studies on a certain topic. After a systematic search and retrieval of relevant studies, the results are scaled to a common unit of measurement, expressed as effect sizes, usually *d* (Cohen 1988) and allowance is made for different sources of error, in particular, sampling error. The assumption of the meta-analytic approach is that these disattenuated effect sizes are all estimates of a common effect that underlies a whole population of studies. Where variation in effect sizes emerges that is not due to sampling error, the analysis seeks to explain those differences in terms of variation arising from the different contexts and characteristics of the primary studies. As a result of this process, meta-analysis allows the: (a) estimation of effect size parameters, (b) explanation of differences in estimates of effect size, (c) examination of stronger estimates of effect sizes in particular situations, and (d) modelling of factors producing effects in different contexts and under different conditions.

Method

It has been argued (e.g. Cook et al. 1992) that meta-analyses frequently suffered from a lack of transparency with regard to the inclusion or exclusion of primary studies. In order to increase transparency, a summary of the principles guiding the selection of primary studies whose results entered the current meta-analysis is given in Table 1.

Authors have differed in their views on which primary studies to include in a meta-analysis. Slavin (1984, 1986), for example, argued that only primary studies of sound methodological quality should be included in a meta-analysis. Glass et al. (1981), on the other hand, claimed that the breadth of the available evidence should be used when synthesising the current state of knowledge in a particular research area. This view was also supported by Kulik and Kulik (1989) who argued that meta-analyses with a high quality approach to selecting primary studies were often left with too few studies to allow the statistical analysis of the results.

It should be noted that over and above the criteria given in Table 1, no further evaluation of studies was undertaken to determine the inclusion or exclusion of studies entering the current meta-analysis.

In Appendix 1 an overview of the studies included in this meta-analysis is provided whereby national studies or authors analysing data from national studies are listed first, followed by international assessment programs. After the sequential study number in Column 1, the name of the study or the name of the author who reported the study is listed in the second column and followed by information about the country in which the study was conducted in the third column.

The data that are used in the meta-analysis are provided in Columns 4 to 8. The first of these columns contains the effect size in the form of Cohen's *d*. Effect size (ES) is defined by Cohen (1988, p. 8) as follows:

> …it is convenient to use the phrase "effect size" to mean "the degree to which the phenomenon is present in the population", or "the degree to which the null hypothesis is false". Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero.

The reason for Cohen's emphasis on effect sizes stemmed from his criticism of the widespread use of significance tests. Cohen pointed out that the reliance on such tests was misleading not only in that a number of assumptions underlying these tests were frequently not met but in that these tests also provided less information than was possible. While a significance test provided information only as to whether or not the null hypothesis was false, the effect size provided additional information regarding the specific degree to which the hypothesis was false.

Table 1: What the meta-analysis is (not) about

Qualifier	Not about	About
English, verbal ability	Studies that used Grade in English or only general verbal ability as a measure.	Studies had to include some measure of reading comprehension or reading achievement in the language of instruction.
Academic achievement	Studies that did not separate out different aspects of academic achievement – and for example combined mathematics and reading in a single outcome measure were not included.	Studies had to include some measure of reading comprehension or reading achievement in the language of instruction.
Language	Not reading as part of foreign language learning.	Focus was on mother-tongue reading or reading in the languages of instruction.
Information provided	Policy papers, discussion/opinion papers, narrative reviews.	Studies had to provide some data amenable to meta-analysis (means, correlations, regression/path coefficients).
Level of schooling	Primary school level.	Secondary school level (i.e. Grade 6 or 12-year-old students to Grade 12 or 18- year-old students).
Type of variable	Studies that used reading as a predictor, mediator or moderator.	Studies that used reading achievement or reading comprehension as the outcome variable or which focused on correlating various factors or variables with reading achievement.
Reading dimension	Comprehension of a specific type of text or using reading for a specific purpose (e.g. RL's 'documents', 'expository', 'narrative' domains or PISA's 'retrieving', 'interpreting' and 'reflecting' and 'evaluation' skills).	An overall score of performance in reading.
Type of student	Samples that focused on students with disabilities, ethnic minority students.	Samples that were representative of mainstream secondary school students.
Level of data collection	If teacher ratings of student achievement were used; analyses reported at school level (e.g. headmaster studies).	Studies had to focus on student-level variables. Information provided by students.
Type of information	If results were not separated in studies of primary and secondary school students.	Information on effect sizes (e.g. correlation coefficient or mean differences) had to be reported for secondary school students).
Type of publication	Dissertations.	Journal articles (as retrieved from a search using 'secondary' and 'student factors' and 'reading achievement' or 'reading performance' in Eric, Web of Science and PsycINFO and selected according to the criteria in this table) or published study reports.
Date of study	Prior to 1970 or after 2002.	1970-2002

> Thus whether measured in one unit or another, whether expressed as a difference between two population parameters or the departure of a population parameter from a constant or in any other suitable way, the ES can itself be treated as a parameter which takes the value zero when the null hypothesis is true and some other specific nonzero value when the null hypothesis is false, and in this way the ES serves as an index of degree of departure from the null hypothesis. (Cohen, 1988, p. 10)

The way in which to interpret the effect size of Cohen's *d* is as follows. If *d* is calculated to be 0.2, then the means differ by two-tenths of a standard deviation. According to Cohen (1988, p.21) *d* is a pure number, which is freed of dependence upon any specific unit of measurement. A value of 2.0 for *d* indicates that the means differ by two standard deviations. An examination of the effect sizes in the third column of Appendix 1 reveals that values range from –0.87 (Study 57), indicating higher achievement of male students, through 0.00 (Studies 106, 143, 144), indicating

no gender differences in reading achievement, to 0.59 (Study 86), indicating a higher performance by female students by about six-tenths of a standard deviation.

The column that follows the effect size is labelled '*v*' which is the squared standard error of *d* (Raudenbush and Bryk 1985; for further details on how '*v*' was calculated, see Equations 2 and 3 below). In the next column, a '1' is assigned if the reading test was administered in English to the whole or the majority of the sample and a '0' if the test was administered in a language other than English. Through the inclusion of this variable in the analysis, it is intended to investigate the potential impact of whether or not the test is administered in English on the variation in gender differences in reading. This is particularly interesting for those assessment programs in which test design takes place in English while tests are administered in many different languages (i.e. PISA, RC, RL). In Column 7, information regarding the mean age of the sample for each study is recorded in order to examine whether or not the possible gender gap in reading increases or decreases with age.

The next column is labelled 'time' and indicates whether a study was undertaken prior to or after 1991. Thus, results from the Reading Literacy Study were assigned a '0' as it was conducted in 1990-91 whereas data provided by the PISA-2000 assessment were assigned a '1' as they had been collected in and after 1992. The reason for choosing 1991 as a cut-off point was the fact that it was only after that date that many testing programs started to use procedures for eliminating at least in part the effects of measurement error from the estimated scores (see Adams, 2005; Wu, 2005) as well as using plausible values in their reports and analyses. Thus, this dummy variable was generated to allow for the examination of possible effects stemming from the way in which reading scores were calculated.

COMMENT ON PARTICULAR MAJOR STUDIES

Below, a short description is given of the assessment programs from which most of the primary study results in the meta-analysis are taken, including information regarding the way in which reading scores were calculated in each program.

Reading Comprehension Study

The first large-scale cross-national survey of reading was conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 15 education systems as the Reading Comprehension Study which formed part of IEA's Six Subject Survey in 1970-71. The reading comprehension test consisted of eight passages and 52 multiple-choice test items that were designed to measure four categories, namely the ability to: (a) follow the organisation of a passage; (b) respond to questions that were specifically answered in the passage; (c) draw inferences from a passage; and (d) identify the writer's purpose. Items were administered to a representative sample of 14-year-old students in each of the participating education systems (Thorndike 1973). In all analyses, Thorndike (1973) used test scores corrected for guessing as indicators of reading performance. These were also the scores used in the current meta-analysis.

Reading Literacy Study

The Reading Literacy Study was the next study of reading performance conducted by IEA in 1990-91. This time, 31 education systems participated at the 14-year-old level (Population B). As in the first study, samples representative of the target population were drawn in each country under the supervision of an international sampling referee. The design of the reading test had shifted from an emphasis on skills to an emphasis on different types of reading materials, namely narrative, expository and documents. As a consequence, students had to answer a total of 89 multiple-choice items relating to 19 passages (Elley 1994). Reading scores based on the one-

parameter model developed by Rasch (1960) were calculated as indicators of performance in reading, whereby one overall reading score was calculated as well as three separate ones, one for each domain. While most of the reporting was undertaken by domain, the score used in the current meta-analysis is the overall score for male and female students from Population B for each country that participated in the study (Elley, 1994, p.106).

Programme for International Student Assessment (PISA)

In the late twentieth century, the Organisation for Economic Co-operation and Development (OECD) launched its Programme for International Student Assessment (PISA) with the main aim to compare the performance of students towards the end of compulsory schooling in key subject areas, namely Mathematics, Reading and Science across its member countries. The focus of the first round of data collection in 2000, in which a total of 43 OECD and non-OECD member countries participated, was on reading. The reading test assessed performance on five processes, namely: (a) retrieving information, (b) forming a broad general understanding, (c) developing an interpretation, (d) reflecting on and evaluating the content of a text, and (e) reflecting on and evaluating the form of a text. Items were of the multiple choice as well as the open constructed-response type and related to continuous and non-continuous texts. Each participating country had to survey a nationally representative sample of 15-year-old students and comply with the sampling guidelines of the OECD (Adams and Wu 2002).

In PISA-2000, two types of reading scores were calculated, namely Warm's (1985) weighted likelihood estimator (WLE) and Bayesian estimation procedures with plausible values (PV) (Adams and Wu 2002). While the weighted likelihood estimator uses the actual score a student obtained as the most likely, plausible values are random numbers that are…

> […] drawn from a distribution of scores that could be reasonably assigned to each individual-that is, the marginal posterior distribution. As such, plausible values contain random error variance components and are not optimal as scores for individuals. Plausible values as a set are better suited to describing the performance of the population. (Adams and Wu 2002, p. 107)

For the international PISA-2000 data set, six WLEs were calculated for each student, one for each of the subject areas tested, namely mathematics, reading and science and three for the reading sub-scales, namely retrieving information, interpreting texts and reflection and evaluation. In addition, 30 plausible values were generated for each student: five for each of the three subject areas and five for the three reading sub-scales. The country-level average scores used in this meta-analysis were the first plausible value mean score (PV1read) for male and female students for the overall reading scale, weighted by the population student weight (w_fstuwt)[1].

NAEP Studies

The National Assessment of Educational Progress (NAEP) is an assessment program run by the National Center for Education Statistics (NCES) in the United States Department of Education.

---

[1] The PISA 2000 technical report (Adams and Wu, 2002) recommends the application of the student weight (w_ fstuwt) for all between country-analyses such as the application in this meta-analysis. The report also recommends that ideally, analyses should be repeated for each of the five plausible value estimates. This was not done in the current analysis which used the first plausible value (PV1Read) only. To illustrate how close the population estimates for plausible values are, an example is given from the German PISA 2000 data set.
For girls (all weighted by student population weight): PV1Read=501.9074;
PV2Read=502.2901; PV3Read=502.2903; PV4Read=502.4483; PV5Read=502.0534.
For boys (all weighted by student population weight): PV1Read=467.7509;
PV2Read=468.7154; PV3Read=467.0083; PV4Read=467.9008; PV5Read=466.3843.

Since 1969, NAEP has conducted studies in a number of subject areas, including reading, to assess achievement levels of nationally representative student samples in Grades 4, 8, and 12. In the most recent reading test design, students were assessed on four aspects of reading. These covered the: (a) forming a general understanding; (b) developing interpretation; (c) relating information in the text to own knowledge and experience; and (d) examining content and structure, which required critical evaluation and an appreciation of the effects of text features such as irony, humour and organisation. To this end, the reading comprehension test employed multiple-choice questions, designed to test students' understanding of individual texts, as well as their ability to integrate and synthesise ideas across the texts and constructed-response questions, which required students to construct their own answers (Plisko 2003).

Over the more than 35 years that NAEP has been the so-called 'Nation's report card' in the United States, the way in which reading scores were calculated has changed as NAEP has used Bayesian estimation procedures and plausible values for its more recent assessment programs (see Beaton 1987; Campbell et al. 2000; Gorman 2005). Thus, the data employed in the current meta-analysis from NAEP assessments between 1971 and 1980 used scores corrected for guessing while the assessments between 1992 and 2003 used plausible values and weighted likelihood estimates.

Australian Studies

In Australia, data on the reading performance of secondary school students were available from a number of studies. They included the 1975 and 1980 studies Australian Studies in School Performance (ASSP) and Australian Studies in Student Performance (1980), the Youth in Transition Study (YIT) in 1989, and the Longitudinal Surveys of Australian Youth (LSAY) that were conducted in 1995 and 1998. The ASSP data included national samples at both ages 10 and 14 years, whereas the Youth in Transition Study and the longitudinal surveys collected data from 14-year-olds only (Rothman 2002).

The reading tests used in these various studies were not the same. The 1975 test was designed to assess minimum competency, and therefore focused on the lower levels of achievement, while the later tests generally covered a wider range of student performance. However, all tests contained a number of common items, which were used in the analysis of trends in reading achievement over time (Marks and Ainley 1996).

The Monitoring Standards in Education (MSE) program in Western Australia started with the Random Sample assessment program in 1990 with data collections that occurred in 1992, 1995, 1997, 1999 and 2001 whereby ten per cent of students in each of Grades 3, 7, and 10 were tested. In 1998, the Western Australian Literacy and Numeracy Assessment (WALNA) population testing began with Grade 3 students. Subsequently, the assessment of Grade 5 was introduced and the Grade 7 was also included. Data collection from Grade 10 students has continued to be undertaken as part of the Random Sample assessment program. Reading performance was assessed on a range of texts that included continuous texts, for example poems, media releases, narrative extracts, as well as non-continuous texts such as charts or tables.

COMMENT ON STATISTICAL PROCEDURES EMPLOYED

It might be argued that the focus of the current meta-analysis on gender differences in reading achievement at the secondary school level was sufficiently narrow to allow for a relatively straight-forward investigation. Unfortunately, this was not the case. Studies that were retrieved as a result of the literature search differed markedly not only in design, sample size, scope and the scale of the reading score but also in the reporting of results. Thus, results were frequently not reported in terms of standardised effect sizes but in terms of correlation coefficients, regression

coefficients from single-level and multi-level analyses, sums of squares, percentage differences or mean differences. Hence, some form of standardisation of the results reported by the different studies was required in order to arrive at a metric-free effect size (ES) that could be processed further in the meta-analysis. The formulae that were employed in the conversion of correlation coefficients, standardised scores, and proportions of test items answered correctly to standardised effect sizes are given in Appendix 2.

As the next step, a so-called 'v-known' hierarchical linear model analysis (Raudenbush et al. 2001, Hox 1995) was undertaken. V-known models may be considered a special case of a two-level hierarchical linear model. In general, hierarchical linear models seek to take into consideration the nested structure of many data sets whereby, for example, students (Level-1) are nested within schools (Level-2). In these instances, variation in the outcome variable at Level-1, frequently a measure of student performance in some subject area, is sought to be explained by variables at Level-1, for example, Gender or Socio-economic status or Homework effort as well as by variables at Level-2, for example, School resources, Size of school, or Location of school. In a meta-analysis the hierarchical structure of the data is such that the within-study variation is modelled at Level-1 while between-study variation is used at Level-2 to explain variability at Level-1. In other words, multilevel modelling as applied to meta-analysis proceeds in two steps. First, it examines whether the within-study results at Level-1 are homogeneous or heterogeneous. If results are homogeneous, the effect sizes may be combined into one average outcome. If the results are heterogeneous, between-study characteristics such as Type of study design or Type of study participants are examined at Level-2 to see whether or not they contribute to explaining differences in results. The reason why Raudenbush and Bryk (2002) labelled these multilevel models for meta-analysis 'v-known models' stems from the fact that the variability at Level-1 is considered to be sampling variability which is known if the relevant sampling distribution and sample sizes are known. Below, the v-known HLM meta-analysis is worked through for the current meta-analysis based on the considerations put forward by Raudenbush and Bryk (2002, p. 208-210).

The effect size (ES) estimate, d_j, for most of the studies listed in Appendix 1 is the standardised mean difference between the average reading scores for female and male students:

$$d_j = \left(\bar{Y}_{Ej} - \bar{Y}_{Cj}\right) / S_j \qquad [1]$$

where

$\bar{Y}_{Ej}$ is the average reading score for the experimental group, that is, female students;

$\bar{Y}_{Cj}$ is the average reading score for the control group, that is, male students;

S_j is the pooled, within-group standard deviation.

Each of the effect sizes recorded in Appendix 1 is an estimate of the population mean difference between the experimental group, which in this context, consists of female students and the control group, which, in this instance, is male students. Thus, in the second study in the Appendix 1, for example, female students score one-tenth of a standard deviation higher than male students.

With reference to Hedges (1981), Raudenbush and Bryk (2002) stated that d_j follows a normal distribution with variance V_j where

$$V_j = (n_{Ej} + n_{Cj}) / (n_{Ej}\, n_{Cj}) + \delta_j{}^2 / [2(n_{Ej} + n_{Ej})] \qquad [2]$$

and assert that "it is common to substitute d_j for δ_j and then assume that V_j is "known"" (Raudenbush and Bryk, 2002, p. 209). While the above formula applies to instances where effect sizes are calculated on the basis of mean differences, the following formula applies to effect sizes calculated on the basis of correlation coefficients:

$$V_j = 1/(n_j - 3) \quad [3]$$

In order to define the hierarchical model for meta-analytic problems, equations have to be formulated at two levels. The model at Level-1, that is the within-study model, is:

$$d_{ij} = \delta_j + e_{ij} \quad [4]$$

where each of the effect sizes for the 147 studies in the current meta-analysis is considered to be one estimate of the underlying population parameter δ_j plus the sampling error associated with each estimate, e_{ij} with $e_{ij} \sim N(0, V_j)$ (where i is the within study subsample, and j is the study sample).

At Level-2, study characteristics and random error are considered to predict the unknown effect size δ_j. Thus, the model at Level-2, that is the between-study model, is:

$$\delta_j = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \gamma_3 W_{3j} + u_j \quad [5]$$

Where: $W_{1j}, ..., W_{3j}$ = are the study characteristics, namely:

(a) Two general predictor variables:

W_1 = English as the language of test administration, W_2 = Age.

(b) Whether a study was conducted up to and including 1991 or from 1992 onwards: W_3 = Time,

$\gamma_0, ..., \gamma_3$ are the regression coefficients associated with the study characteristics W_1 to W_3,

u_j is Level-2 random error where $u_j \sim N(0, \tau)$.

In order to combine the two-levels into a single model, δ_j in Equation 4 has to be replaced by δ_j from Equation 5:

$$d_{ij} = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \gamma_3 W_{sj} + u_j + e_{ij} \quad [6]$$

In summary, the Level-1 outcome variable in the meta-analysis is the effect size which quantifies the difference between male and female students' performance in reading reported by each study. In case the variation in effect sizes is found not to be due to chance, the analysis reveals the extent to which variables W_1 to W_3 contribute to explaining the variance.

W_1 and W_2 are specified to examine the potential effects of two variables, namely whether or not English is the language of testing and the average age of the students in a particular study. This allows the examination of two questions. First, since most of the instrument construction for international tests is undertaken in English and with an interest in gender equitable materials in that language, gender differences may be less pronounced in countries where English is the language of instruction and test administration. Second, as male students mature later than female students and reading is basically a process of reasoning (Lietz 1996; Thorndike 1917), gender differences may decrease with increasing age.

The effect sizes used in this meta-analysis were taken from large-scale national and international studies. Thus, in order to examine possible systematic impact on effect sizes of the way in which scaled performance scores were calculated from 1992 onwards, the dummy variable W_3 (Time) was created to indicate whether a study was undertaken up to and including 1991 (dummy code '0') or from 1992 onwards (dummy code '1'). In this way, it was possible to investigate whether or not any systematic difference, associated with the time period in which a study was conducted, emerged. Evidence supporting the introduction of such a time variable is given in the section below entitled 'Some problems involved in comparing effect sizes from different testing programs'.

RESULTS

As noted above, the first step in a meta-analysis using HLM was to examine whether the effect sizes from the different primary studies are homogenous or heterogeneous. In the case where heterogeneity could be ascertained, an analysis was undertaken to investigate the way in which possible study characteristics could contribute to the variability in effect sizes.

Testing the Null Model

It can be seen in Table 2 that the estimated grand-mean effect size, the intercept in the model, is positive and small, γ_{10} ($G10$) =0.18, which means that, on average, female secondary students performed about 0.18 standard deviation units above male secondary students. It should be noted that the number of degrees of freedom is 146, one fewer than the number of studies in the analysis, as one degree of freedom is needed for the estimation of the unconditional model. The only parameter to be estimated is the intercept.

Furthermore, the estimated variance of the effect parameter is 0.024 with a standard deviation of 0.15 indicating important variability in the effect sizes. Moreover, the Chi-square value (2557.46) and corresponding p-value (0.000) confirm that this variance is not due to chance and that the residual variance is significantly different from zero. As a consequence, the analysis can proceed to examine which of the predictor variables that reflect study characteristics are able to explain this variance.

Table 2: Final estimation of fixed effects: Unconditional 'v-known' model

Standard Approx. Fixed Effect	Coefficient	Error	T-ratio	d.f.	P-value
For EFFSIZE, B1					
INTRCPT2, G10	0.184245	0.014021	13.141	146	0.000

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
EFFSIZE, U1	0.15420	0.02378	146	2557.46428	0.000

Statistics for current covariance components model

Deviance = 964.115997 df = 2

Some Problems Involved in Comparing Effect Sizes from Different Testing Programs

While all the testing programs under consideration in this article are concerned with gender differences in achievement, the present information is associated with comparisons that are obtained in many different ways, which are discussed in some detail in Appendix 2. Moreover, the

testing programs most probably calculated their sampling variance estimates in different ways in attempting to take into consideration the hierarchical structure of the sample data. Efforts made to develop a set of procedures for this study in order to achieve uniformity in the calculation of effect sizes and estimates of '*v*' proved to be not only frustrating but also unrewarding. Consequently, it was assumed in an earlier analysis (Lietz, in press), that, in general, a particular testing program would use a common procedure across the different studies over time, and allowance could be made for a treatment effect for each of the different testing programs by including dummy variables indicating whether a study belonged to the Reading Comprehension, the Reading Literacy, the PISA, the NEAP or the Australian Testing Program. Likewise, two dummy variables were included in the analysis to indicate the two main bases for estimating effect sizes, namely means and correlations and the corresponding procedure to estimate '*v*'. In Table 3 the results of this earlier analysis (Lietz, in press) are recorded. These analyses included the aforementioned dummy variables plus whether or not English was the language of testing and age as Level-2 predictors. In the table, regression coefficients, their standard errors, t-values associated with each of these predictors as well as the approximate degrees of freedom and p-values that were obtained in initial analyses of the data (Lietz, in press) are presented.

Results showed only small difference between the effect sizes calculated from means (ESMEAN G18 = 0.160) and effect sizes calculated from correlations (ESCORR G19 = 0.188). In contrast, differences between the estimates of the effect sizes for the Reading Comprehension Study (RC G13 = -0.076), the Reading Literacy Study (RL G14 = 0.017) and PISA (PISA G15=0.235) were substantially large. This evidence suggested that the way in which the variance estimates employed in the different methods of estimating effects sizes warranted closer attention.

It was the substantial differences between the coefficients for the different testing programs shown in Table 3 that led to a re-examination of the effect size data. In particular, it became interesting to examine whether the differences may not be so much stemming from the different testing programs per se but be a consequence of different procedures for calculating test scores that were introduced in the early 1990s.

Table 3: Final estimation of fixed effects: 'v-known' model with all predictors included

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For EFFSIZE, B1					
INTRCPT2, G10	-0.129615	0.162295	-0.799	137	0.425
ENGLISH, G11	0.021762	0.029647	0.734	137	0.463
AGE, G12	0.000490	0.009474	0.052	137	0.959
RC, G13	-0.076360	0.059346	-1.287	137	0.198
RL, G14	0.016580	0.040229	0.412	137	0.680
PISA, G15	0.235441	0.038756	6.075	137	0.000
NEAP, G16	0.181077	0.047319	3.827	137	0.000
OZ, G17	0.206835	0.039553	5.229	137	0.000
ESMEAN, G18	0.159932	0.059703	2.679	137	0.008
ESCORR, G19	0.187569	0.079254	2.367	137	0.018

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
EFFSIZE, U1	0.09866	0.00973	137	1050.40170	0.000

Statistics for current covariance components model

Deviance = 882.930946 df = 2

In order to summarise the problems raised in this section, it is recognised that in this article the author is attempting to bring together in a meta-analysis the results obtained from the calculation of effect sizes and estimates of '*v*' using very different and perhaps in certain cases possibly

inappropriate procedures. Results of the earlier analysis presented in Table 3 showed that these different procedures could possibly be allowed for through the use of dummy variables for the different testing programs. Nevertheless, what is clear is that the comments being made in informal discussions about changes in gender differences in levels of reading performance, being due to changes in reading habits between boys and girls, and the effects of watching TV or working at computers are not warranted until more work is undertaken to examine the procedures used in the different studies that have been undertaken over time. Hence, the following section reports results of a meta-analysis which includes time as a predictor at Level 2.

Change in Recorded Effect Sizes Over Time

In order to examine the potential effect of time on the extent of gender differences, a HLM model which includes the predictors specified in Equations 5 and 6 above was examined and the results are presented in Table 4. Note that the degrees of freedom are now reduced to 143 as, in addition to the intercept, three potential Level-2 predictors, namely Age, whether or not English was the language of testing and Time, needed to be estimated.

Of the three possible predictors only one emerges with a significant effect whereas the remaining two do not contribute to explaining the variability in effects sizes. Thus, Age and whether or not English (ENG) was the language of test-administration do not emerge as significant predictors of gender differences. In other words, the gender gap does not decrease with age, which may have supported the maturational viewpoint whereby reading comprehension is also a function of maturity and, since boys mature at a later age, differences between boys' and girls' reading performance may decrease with increasing age. Likewise, there is no evidence to suggest that gender differences are more or less pronounced in countries where English is not the language of test administration.

However, the impact of the variable Time on the effect size is positive $\gamma_{13}G(13) = 0.24$ and highly significant ($p=0.00$). The way in which this variable is coded means that studies prior and up to 1991 receive the lower ('0') code while studies from 1992 onwards are assigned the higher ('1') code. As a consequence, because the effect of this variable is estimated to involve a gender difference in favour of girls of about 0.24 units higher for studies that had been conducted since 1992 than for those studies that were undertaken prior to that year.

Table 4: Final estimation of fixed effects: 'v-known' model with predictors included

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For EFFSIZE, B1					
INTRCPT2, G10	0.100440	0.113192	0.887	143	0.375
ENG, G11	-0.017951	0.018880	-0.951	143	0.342
AGE, G12	-0.001950	0.007544	-0.258	143	0.796
TIME, G13	0.243677	0.018519	13.159	143	0.000

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
EFFSIZE, U1	0.08786	0.00772	143	984.49931	0.000

Statistics for current covariance components model

Deviance = 776.796393 df = 2

In order to arrive at the final hierarchical model, the two between-study variables that did not contribute significantly to explaining differences in effect sizes, namely English and Age were

removed from the model. Results of the final model in which only the variable Time is included as a predictor are shown in Table 5.

The intercept in Table 5 is positive and small (0.06), and not significantly different from zero. This finding, in addition to the contrasting results for the intercepts presented in Tables 2, 3 and 4, provides evidence that given the data in this analysis male students performed at a slightly lower level in reading than did female students. However, a sizable and significant positive effect is recorded for Time $\gamma_{11} G(11) = 0.25$ which indicates that since 1992 girls outperformed boys to a considerably greater extent when compared with studies up to and including 1991.

A comparison of the deviance values allows an evaluation of the three models under review, namely the unconditional model, the model which includes all three predictors and the final model with only time as a predictor. Thus, the deviance which is highest for the unconditional model with a value of 964.1 is reduced to 776.8 for the second model. For the final model, in turn, the deviance is further reduced to a value of 743.2 which indicates that the last model provides the best fit to the data, and the removal of the non-significant variables of Age and English yield a better fitting model to the data.

Table 5: Final estimation of fixed effects: 'v-known' model with 'Time' as a predictor

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For EFFSIZE, B1					
INTRCPT2, G10	0.059692	0.013339	4.475	145	0.000
TIME, G11	0.247168	0.018041	13.700	145	0.000

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
EFFSIZE, U1	0.08751	0.00766	145	1043.16574	0.000

Statistics for current covariance components model

Deviance = 743.151878 df = 2

The variance estimates of the unconditional model (0.02378) and the final model (0.00766) can be used to calculate the proportion of variance explained in study results. Thus, the final v-known model explains 67.8 per cent ((0.02378-0.00766)/0.02378) of the variance in the data. Complementary information is provided by the chi-square (1043.17) and p-values (0.000) computed for the estimated variance of the effect parameters in the final model of 0.007 which corresponds to a standard deviation of 0.087 and indicates that important variability still exists in the effect sizes. Thus, while the between-study variable Time included in the final v-known model explains about two-thirds of the differences in effect sizes a moderate amount of variability remains to be explained by factors other than those included in this analysis.

CONCLUSIONS

In this study, a meta-analysis of large-scale studies between 1970 and 2002 in the area of reading achievement at the secondary school level with a focus on gender differences was conducted. The meta-analysis was conceptualised as a special application of a two-level hierarchical linear model whereby in a first step, it was examined whether the effect sizes differed more than could be expected due to sampling error. Once results had been ascertained to be sufficiently heterogeneous, characteristics at Level-2 were examined and the way in which they could explain differences between effect sizes at Level-1. Level-2 variables included in the hierarchical linear

model covered the age of study participants, and whether or not a study was conducted in a country where English was the language of test administration. In addition, because of the results from an initial meta-analysis which suggested that gender effects were more pronounced in more recent assessment programs a variable indicating whether studies had been conducted prior to or after 1992, was introduced into the analyses.

It is seen that (a) gender differences exist across the 147 studies under review that are not due to chance; and (b) about two-thirds of the variance associated with these differences can be explained by the introduction of a Time variable into the meta-analysis.

Thus, the gender gap in favour of girls is even more pronounced for the assessment programs that have been conducted since 1992. Possible explanations for the origins of these greater differences could be related to item selection procedures or contextual changes surrounding reading in society. Such explanations would appear unlikely, given the stringent psychometric procedures to investigate item bias, in particular with respect to Gender, that had been employed in the large reading assessment programs under review. Likewise, there was little evidence of a general decline in societal support for reading aimed particularly at boys since 1992. Thus, it might be a reasonable explanation that the increase in gender differences for more recent assessment programs might stem from changes in the way in which performance were calculated prior to and after 1992. More specifically, the change to using Bayesian estimation procedures and plausible values or weighted likelihood estimates might have introduced some systematic bias into the effect size indexes as a consequence of a reduction in the within group variance. Alternatively, it might be argued that either prior to 1992 or after 1992 the estimates made of gender differences in reading achievement were basically wrong, because inappropriate estimates of between group variance were being employed in the calculation of effect sizes. Consequently, any discussion of change over time in gender differences in reading achievement and possibly other aspects of educational performance would be inappropriate until the issues raised in this article are resolved.

APPENDIX 1:
STUDIES IN THE META-ANALYSIS
IN ALPHABETICAL ORDER OF AUTHOR OR STUDY

No.	Study/Author	Country	ES(*d*)	*v*	English	Mean Age	Time
1	ASSP 1975	Australia	0.090	0.001	1	14.00	0
2	ASSP 1980	Australia	0.110	0.001	1	14.00	0
3	YIT 1989	Australia	0.080	0.001	1	14.00	0
4	LSAY 1995	Australia	0.190	0.000	1	14.00	1
5	LSAY 1998	Australia	0.230	0.000	1	14.00	1
6	WA monitoring 1992	Australia	0.313	0.003	1	12.00	1
7		Australia	0.344	0.003	1	15.00	1
8	WA monitoring 1995	Australia	0.344	0.003	1	12.00	1
9		Australia	0.389	0.003	1	15.00	1
10	WA monitoring 1997	Australia	0.193	0.003	1	12.00	1
11		Australia	0.448	0.003	1	15.00	1
12	WA monitoring 1999	Australia	0.406	0.003	1	12.00	1
13		Australia	0.434	0.003	1	15.00	1
14	WA monitoring 2001	Australia	0.306	0.000	1	12.00	1
15		Australia	0.496	0.004	1	15.00	1
16	WA monitoring 2002	Australia	0.230	0.000	1	12.00	1
17	Fuller et al. 1994	Botswana	0.143	0.000	1	15.00	1
18		Botswana	0.192	0.000	1	16.00	1
19	GambellandHunter2000	Canada	0.237	0.042	1	13.00	1

20		Canada	0.247	0.042	1	16.00	1
21	Glossop et al. 1979	England	-0.155	0.006	1	15.00	0
22	Gorman et al. 1982	Engl., Wales, Nth. Ireland	0.013	0.001	1	15.75	0
23		Nth. England	0.014	0.004	1	15.00	0
24		Midlands	-0.040	0.005	1	15.00	0
25		Sth. England	0.025	0.003	1	15.00	0
26		Wales	-0.023	0.005	1	15.00	0
27		Nth. Ireland	0.136	0.004	1	15.00	0
28	Youngman 1980	UK	0.040	0.003	1	12.00	0
29		UK	0.283	0.003	1	12.00	0
30	Hogrebe et al 1985	USA,HSB-80	-0.050	0.000	1	17.00	0
31		USA,HSB-80	-0.090	0.000	1	15.00	0
32	LevineandOrnstein 1983	USA,NAEP-71	0.056	0.005	1	13.00	0
33		USA,NAEP-71	0.048	0.005	1	17.00	0
34		USA,NAEP-75	0.056	0.005	1	13.00	0
35		USA, NAEP-75	0.040	0.005	1	17.00	0
36		USA, NAEP-80	0.048	0.005	1	13.00	0
37		USA, NAEP-80	0.038	0.005	1	17.00	0
38	NAEP 2003	USA	0.220	0.005	1	13.00	1
39	NAEP 2002	USA	0.180	0.005	1	13.00	1
40	NAEP 1998	USA	0.280	0.005	1	13.00	1
41	NAEP 1994	USA	0.300	0.005	1	13.00	1
42	NAEP 1992	USA	0.260	0.005	1	13.00	1
43	NAEP 2002	USA	0.320	0.005	1	17.00	1
44	NAEP 1998	USA	0.320	0.005	1	17.00	1
45	NAEP 1994	USA	0.280	0.005	1	17.00	1
46	NAEP 1992	USA	0.200	0.005	1	17.00	1
47	NeumanandProwda 1982	Connecticut 1978-79, USA	0.120	0.000	1	13.00	0
48		Connecticut 1978-79, USA	0.100	0.000	1	16.00	0
49	HedgesandNowell1995	USA, NELS-88	0.090	0.005	1	13.00	0
50		USA, NLS-72	0.050	0.005	1	17.00	0
51		USA, NLSY-80	0.180	0.005	1	18.50	0
52	OaklandandStern1989	Texas, USA	0.006	0.003	0	10.50	0
53	Project Talent 1960	USA	0.150	0.005	1	15.00	0
54	ShillingandLynch 1985	Pennsylvania, USA	0.161	0.005	1	13.00	0
55	Johnson 1973-74	Canada	0.172	0.041	1	12.00	0
56		England	-0.250	0.039	1	12.00	0
57		Nigeria	-0.870	0.038	1	13.00	0
58		USA	0.103	0.041	1	12.00	0
59	PISA2000	Australia	0.330	0.001	1	15.00	1
60	PISA2000	Austria	0.250	0.001	0	15.00	1
61	PISA2000	Belgium	0.330	0.001	0	15.00	1
62	PISA2000	Canada	0.320	0.000	1	15.00	1
63	PISA2000	Czech Republic	0.370	0.001	0	15.00	1
64	PISA2000	Denmark	0.250	0.001	0	15.00	1
65	PISA2000	Finland	0.510	0.001	0	15.00	1
66	PISA2000	France	0.290	0.001	0	15.00	1
67	PISA2000	Germany	0.340	0.000	0	15.00	1
68	PISA2000	Greece	0.370	0.003	0	15.00	1
69	PISA2000	Hungary	0.310	0.002	0	15.00	1
70	PISA2000	Iceland	0.400	0.000	0	15.00	1
71	PISA2000	Ireland	0.290	0.001	1	15.00	1
72	PISA2000	Italy	0.380	0.001	0	15.00	1
73	PISA2000	Japan	0.300	0.004	0	15.00	1
74	PISA2000	Korea	0.140	0.001	0	15.00	1
75	PISA2000	Luxembourg	0.270	0.000	0	15.00	1
76	PISA2000	Mexico	0.210	0.001	0	15.00	1

77	PISA2000	New Zealand	0.460	0.001	1	15.00	1
78	PISA2000	Norway	0.430	0.001	0	15.00	1
79	PISA2000	Poland	0.360	0.002	0	15.00	1
80	PISA2000	Portugal	0.240	0.002	0	15.00	1
81	PISA2000	Spain	0.240	0.001	0	15.00	1
82	PISA2000	Sweden	0.370	0.001	0	15.00	1
83	PISA2000	Switzerland	0.300	0.002	0	15.00	1
84	PISA2000	UK	0.250	0.001	1	15.00	1
85	PISA2000	US	0.280	0.005	1	15.00	1
86	PISA2000	Albania	0.590	0.005	0	15.00	1
87	PISA2000	Argentina	0.440	0.005	0	15.00	1
88	PISA2000	Brazil	0.160	0.001	0	15.00	1
89	PISA2000	Bulgaria	0.480	0.005	0	15.00	1
90	PISA2000	Chile	0.250	0.005	0	15.00	1
91	PISA2000	Hong Kong	0.150	0.005	1	15.00	1
92	PISA2000	Indonesia	0.200	0.005	1	15.00	1
93	PISA2000	Israel	0.150	0.005	1	15.00	1
94	PISA2000	Latvia	0.530	0.005	0	15.00	1
95	PISA2000	Liechtenstein	0.320	0.002	0	15.00	1
96	PISA2000	Macedonia	0.510	0.005	0	15.00	1
97	PISA2000	Peru	0.060	0.005	0	15.00	1
98	PISA2000	Romania	0.130	0.005	0	15.00	1
99	PISA2000	Russia	0.380	0.002	0	15.00	1
100	PISA2000	Thailand	0.420	0.005	1	15.00	1
101	PISA2000	Netherlands	0.300	0.001	0	15.00	1
102	RC 1970-71	Belgium(Fl.)	0.100	0.036	0	14.00	1
103	RC	Belgium(Fr.)	0.345	0.056	0	14.00	0
104	RC	Chile	-0.242	0.010	0	14.00	0
105	RC	England	0.201	0.007	1	14.00	0
106	RC	Finland	0.000	0.014	0	14.00	0
107	RC	Hungary	0.040	0.005	0	14.00	0
108	RC	India	0.040	0.007	1	14.00	0
109	RC	Iran	-0.060	0.033	0	14.00	0
110	RC	Israel	-0.060	0.008	1	14.00	0
111	RC	Italy	0.040	0.003	0	14.00	0
112	RC	Netherlands	-0.060	0.021	0	14.00	0
113	RC	New Zealand	0.040	0.014	1	14.00	0
114	RC	Scotland	-0.140	0.015	1	14.00	0
115	RC	Sweden	0.120	0.011	0	14.00	0
116	RC	USA	0.080	0.007	1	14.00	0
117	RL1990-91	Trin and Tobago	0.299	0.011	1	14.40	0
118	RL	Thailand	0.304	0.007	1	15.20	0
119	RL	Ireland	0.284	0.007	1	14.50	0
120	RL	Canada(BC)	0.259	0.005	1	13.90	0
121	RL	Sweden	0.188	0.007	0	14.80	0
122	RL	Finland	0.215	0.015	0	14.70	0
123	RL	Hungary	0.192	0.007	0	14.10	0
124	RL	United States	0.153	0.006	1	15.00	0
125	RL	Iceland	0.167	0.007	0	14.80	0
126	RL	Italy	0.123	0.006	0	14.10	0
127	RL	Netherlands	0.118	0.006	0	14.30	0
128	RL	Cyprus	0.110	0.020	0	14.80	0
129	RL	Germany(E)	0.096	0.010	0	14.40	0
130	RL	Belgium(Fr.)	0.077	0.007	0	14.30	0
131	RL	Botswana	0.140	0.007	1	14.70	0
132	RL	Hong Kong	0.078	0.006	1	15.20	0
133	RL	New Zealand	0.054	0.008	1	15.00	0

134	RL	Philippines	0.077	0.004	1	14.50	0
135	RL	Slovenia	0.079	0.007	0	14.70	0
136	RL	Denmark	0.052	0.005	0	14.80	0
137	RL	Germany(W)	0.051	0.005	0	14.60	0
138	RL	Norway	0.056	0.007	0	14.80	0
139	RL	Spain	0.062	0.003	0	14.20	0
140	RL	Switzerland	0.041	0.003	0	14.90	0
141	RL	Venezuela	0.033	0.006	0	15.50	0
142	RL	Greece	0.015	0.007	0	14.40	0
143	RL	Nigeria	0.000	0.013	1	15.30	0
144	RL	Singapore	0.000	0.007	1	14.40	0
145	RL	France	-0.059	0.008	0	15.40	0
146	RL	Portugal	-0.133	0.008	0	15.60	0
147	RL	Zimbabwe	-0.283	0.007	1	15.50	0

APPENDIX 2:
CALCULATION OF EFFECT SIZES FOR STUDIES IN THE META-ANALYSIS

1. For the Australian studies (reported by Rothman, 2002)

Reported SD of 10. Therefore:

$$d = \frac{\overline{X}_F - \overline{X}_M}{10}$$

2. For studies reporting means and standard deviation for males, means and standard deviation for females and number of cases for each sex (e.g. WA monitoring studies, Hogrebe et al., 1985; Johnson, 1973-74)

$$d = \frac{\overline{X}_F - \overline{X}_M}{\sqrt{\frac{(N_F - 1)s_F^2 + (N_M - 1)s_M^2}{N_F + N_M - 2}}}$$

which is the mean for females minus the mean for males divided by the within-group (also called 'pooled') standard deviation (see Hunter et al., 1982, p. 98).

The reason for using the within-group standard deviation instead of the control-group standard deviation was that the within-group standard deviation had only about half the sampling error of the control-group standard deviation. In addition, Cohen (1988, p. 11) stated that "…the ES index for differences between population means is standardised by division by the common within-population standard deviation."

The reason for subtracting male mean from female mean was that higher average reading performance was expected for females. As a consequence, positive effect sizes denoted superior performance of females whereas negative effect sizes denoted superior performance of males.

3. For the Botswana study (reported by Fuller et al., 1994)

Using t-test values and number of cases to calculate effect size.

First step: Calculate correlation coefficient r from t-test (see Hunter et al., 1982, p. 98):

$$r = \frac{t}{\sqrt{t^2 + N - 2}}$$

Second step: Calculate effect size *d* based on r:

$$d = \frac{r}{\sqrt{1 - r^2 \times p \times q}}$$

where p is the proportion of females and q is the proportion of males in the sample.

Note that Hunter et al. (1982, p. 98) stated that in the case of equal sample sizes for the two groups "[…] for small correlations, this meant *d*=2r[…]".

4. For studies that record percentages (Gambell and Hunter, 2000; and for NAEP 1971, 1975 and 1980 reported by Levine and Ornstein, 1983)

d=SUM(ASIN(p)-ASIN(q))

5. For the United Kingdom study that reported means for females and males plus the respective standard errors and not the standard deviation

Cohen (1988, p. 6) states

> "..one conventional means for assessing the reliability of a statistic is the standard error (SE) of the statistic. If we consider the arithmetic mean of a variable X ($\bar{X}$), its reliability may be estimated by the standard error (SE) of the mean ($SE_{(\bar{X})}$):"

First, obtain SD from SE:

$$SE_{\bar{X}} = \sqrt{\frac{SD^2}{n}}$$

therefore

$$SE_{\bar{X}} = \frac{SD}{\sqrt{n}}$$

therefore

$$SD = SE_{\bar{X}} \times \sqrt{n}$$

Then, replace SD by $SE_{\bar{X}} \times \sqrt{n}$ into the ordinary formula for *d* to calculate the effect size:

$$d = \frac{\bar{X}_F - \bar{X}_M}{\frac{N_F \times SE_F \times \sqrt{N_F} + N_M \times SE_M \times \sqrt{N_M}}{N_F + N_M - 2}}$$

6. For the NAEP studies mean differences (directly off website)

Reported SD of 50, therefore:

$$d = \frac{\bar{X}_F - \bar{X}_M}{50}$$

7. For PISA 2000 studies

Achievement scores were scaled to a mean of 500 and a standard deviation of 100 (Adams and Wu, 2002). Therefore:

$$d = \frac{\bar{X}_F - \bar{X}_M}{100}$$

8. For studies reporting correlation coefficients (includes the Reading Comprehension Study)

$$d = \frac{r}{\sqrt{1 - r^2 \times p \times q}}$$

where p is the proportion of females and q the proportion of males in the sample.

9. For studies reporting partial correlation coefficients, regression weights, or gammas

These were considered more precise estimates of the relationship between gender and reading achievement as the effects of other variables had been partialled out. In other words, these measures provided information on the strength of the relationship between gender and reading achievement after the influences of other variables on the relationships had been taken into account. In line with this argument, betas or gammas of the most complex models were used as a basis for calculating the effect size as these were considered to be better estimates of the relationships between gender and reading achievement, taking into account the other variables.

In line with Pedhazur (1982) regression coefficients could be considered similar in nature to correlation coefficients. Hence, the same formula as for correlation coefficients was used in the calculation of effect sizes from partial correlations, regression coefficients and gammas (from hierarchical linear models).

10. For studies reporting sum of squares as a result of ANOVA analyses (Oakland and Stern, 1989)

The idea that it was legitimate to use the following formula in calculating effect sizes based on sums of squares was put forward by Keppel (1991, p. 437-444).

$$d = \frac{\sqrt{SS_{Sex}}}{\sqrt{SS_{Total}}}$$

Like partial correlation or regression coefficients, this measure was considered to be better as it took into account other variables, such as Race and SES in the analysis by Oakland and Stern (1989).

11. For the Reading Literacy Study

According to Cohen's formula (1988, p. 20):

$$d = \frac{\bar{X}_F - \bar{X}_M}{\sigma}$$

whereby values for means for males and females were taken from Purves and Elley (in Elley 1994, p. 106) and the pooled standard deviation for the overall reading score was taken from Elley and Schleicher (in Elley 1994, p. 57).

REFERENCES

* indicates reports from which data were used in the meta analyses.

Adams, R.J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162-172.

Adams, R., and Wu, M. (Eds.). (2002). *Programme for International Student Assessment (PISA): PISA 2000 Technical Report.* Paris: OECD.

Aiken, L.S. and West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA.: Sage.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-1984 technical report.* Princeton, NJ: Educational Testing Service.

Campbell, J.R., Hombo, C.M. and Mazzeo, J. (2000). *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance, NCES 2000–469*, Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics, NAEP. [Online] http://nces.ed.gov/naep3/pdf/main1999/2000469.pdf [Last accessed 24/11/05].

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, (2nd edn.), Hilldale, NJ.: Lawrence Erlbaum Associates.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Hedges, L.V., Light, R.J., Louis, T.A. and Mosteller, F. (1992). *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Foundation.

Dedze, I. (1995). *Reading Achievement Within the Educational System of Latvia: Results from the IEA Reading Literacy Study*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, April 18-22.

*Elley, W.B. (Ed.). (1994). *The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-Two School Systems*. Oxford: Pergamon Press.

*Fuller, B., Hua, H. and Snyder, C.W. Jr. (1994). Focus on gender and academic achievement. When girls learn more than boys: The influence of time in school and pedagogy in Botswana. *Comparative Education Review*, 38(3), 347-376.

*Gambell, T. and Hunter, D. (2000). Surveying gender differences in Canadian school literacy. *Journal of Curriculum Studies*, 32(5), 689-719

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, 5, 351-379.

Glass, G.V., McGaw, B., and Smith, M.L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, OA.: Sage Publications.

*Glossop, J.A, Appleyard, R., and Roberts, C. (1979). Achievement relative to a measure of general intelligence. *British Journal of Educational Psychology*, 49, 249-257.

*Gorman, T.P., White, J., Orchard, L. and Tate, A. (1982). *Language Performance in Schools. Secondary Survey Report no 1*. Department of Education and Science, Welsh Office, Department of Education for Northern Ireland. London: Her Majesty's Stationery Office.

Gorman, S. (2005). Director for design and analysis, Assessment Division, NCES. Response to the question since when NAEP uses plausible values. Personal e-mail communication via Taslima Rahman 30/11/05.

Hedges, L.V. (1981).Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128.

*Hedges, L.V. and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, 269, 41-45.

Hedges, L.V. and Olkin, I. (1983). Regression models in research synthesis. *The American Statistician*, 37(2), 137-140.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL.: Academic Press.

*Hogrebe, M.C., Nist, S.L. and Newman, I. (1985). Are there gender differences in reading achievement? An investigation using the high school and beyond data. *Journal of Educational Psychology*, 77(6), 716-24.

Hox, J.J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Hunter, J.E. and Schmidt, F.L. (2004). *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*. (2nd ed.) Thousand Oaks, CA: Sage Publications.

Hunter, J.E., Schmidt, F.L. and Jackson, G.B. (1982). *Meta-Analysis. Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage.

Husén, T. (1967). *International Study of Achievement in Mathematics. A Comparison of Twelve Countries. Volume II*. Stockholm: Almqvist and Wiksell.

*Johnson, D.D. (1973-1974). Sex differences in reading across cultures. *Reading Research Quarterly*, 9(1), 67-86.

Johnston, J. and Dunne, M. (1996). Revealing assumptions: Problematising research on gender and mathematics and science education. In L.H. Parker, L.J. Rennie, B.J. Fraser (Eds.) *Gender, science and mathematics. Shortening the shadow* (p. 53-63). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice Hall.

Keeves, J. P. (1988). Sex differences in ability and achievement. In J.P. Keeves (Ed.) *Educational research, methodology, and measurement: An international handbook* (pp. 689-700). Oxford: Pergamon Press.

Kulik, J.A., and Kulik, C.-L.C. (1989). Meta-analysis in education. *International Journal of Educational Research*, 13(3), 221-340.

*Levine, D.U. and Ornstein, A.C. (1983). Sex differences in ability and achievement. *Journal of Research and Development in Education*. 16 (2), 66-72.

Lietz, P. (1996). *Reading Comprehension Across Cultures and Over Time*. Münster/New York: Waxmann.

Lietz, P. (in press). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation*.

Light, R.J. and Smith, P.V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 41(4), 429-471.

Marks, G.N. and Ainley, J. (1996). *Reading Comprehension and Numeracy Among Junior Secondary School Students in Australia. Longitudinal Surveys of Australian Youth, Research Report Number 3*. Melbourne: Australian Council for Educational Research.

*Neuman, S.B. and Prowda, P. (1982). Television viewing and reading achievement. *Journal of Reading*, 25, 666-670.

*Oakland, T., and Stern, W. (1989). Variables associated with reading and math achievement among a heterogeneous group of students. *The Journal of School Psychology*, 27, 127-140

Pedhazur, E.J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*. (2nd ed). Fort Worth, TX: Holt, Rinehart and Winston.

*Plisko, V.W. (2003). *The Release of the National Assessment of Educational Progress (NAEP) The Nation's Report Card: Reading and Mathematics* 2003. [Online] http://nces.ed.gov/commissioner/remarks2003/11_13_2003.asp [Last accessed 18/05/05].

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S.W. and Bryk, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10(2), 75-98.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F. and Congdon, R. (2001). *HLM 5. Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: SSI Scientific Software International.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage Publications.

*Rothman, S. (2002). *Longitudinal Surveys of Australian Youth. Research Report Number 29. Achievement in Literacy and Numeracy by Australian 14-Year-Olds, 1975-1998*. Hawthorne, Vic.: Australian Council for Educational Research. [Online] http://www.acer.edu.au/research/projects/lsay/reports/lsay29.pdf [Last accessed 24/11/05].

*Shilling, F., and Lynch, P.D. (1985). Father versus mother custody and academic achievement of eighth grade children. *Journal of Research and Development in Education*. 18(2), 7-11.

Slavin, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher*, 13, 6-15.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher*, 15, 5-11.

Thorndike, E.L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. Reprinted in *Reading Research Quarterly*, (1971), 6(4), 425-434.

Thorndike, R.L. (1973). *Reading Comprehension Education in Fifteen Countries. International Studies in Evaluation III*. Stockholm: Almqvist and Wiksell.

Tracy, D.M. (1987). Toys, spatial ability, and science and mathematics achievement – Are they related? *Sex Roles*, 17(3-4), 115-138.

*U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, *National Assessment of Educational Progress (NAEP), 2003, 2002, 1998, 1994 and 1992 Reading Assessments. NAEP Data Tool v3.0*. [Online] http://nces.ed.gov/nationsreportcard/naepdata/getdata.asp, search options "subject"= reading; "Grade"=Grade 8 and Grade 12; "State/Jurisdiction"=Nation; Category=Major reporting groups, after pressing "continue" select "gender". [Last accessed 06/06/05].

Wagemaker, H., Taube, K., Munck, I., Kontogiannopoulou-Polydorides, G. and Martin, M. (1996). *Are Girls Better Readers? Gender Differences in Reading Literacy in 32 Countries*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Warm, T.A. (1985). *Weighted Maximum Likelihood Estimation of Ability in Item Response Theory with Tests of Finite Length* (Technical Report CGI-TR-86-08). Oklahoma City: U.S. Coast Guard Institute.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128

*Youngman, M.B. (1980). Some determinant of early secondary school performance. *British Journal of Educational Psychology*, 50, 43-52.

International Education Journal, 2013, Volume 8, 150-159.
ISSN 1443-1475 © 2013

A method for monitoring sub-trends in country-level mathematics achievement on TIMSS

Kelvin Gregory
School of Education, Flinders University *kelvin.gregory@flinders.edu.au*

The Trends in International Mathematics and Science studies provide country-level data for tracking changes in student achievement over time. In this paper the author has developed a method for identifying and monitoring trends in student achievement above or below any specified cut-point on these tests. The method involved the use of the Foster, Greer, and Thorbecke indices, as well as a modified version of these indices. The ability to identify and monitor trends in student achievement at various cut-points on the test should prove useful to policy analysts as well as to governmental and international funding agencies wishing to obtain data on the effectiveness of various programs and policies.

Monitoring trends, achievement, large-scale assessment

INTRODUCTION

Since 1995, the International Association for Evaluation of Educational Achievement (IEA) has conducted three large-scale comparative studies of mathematics and science achievement. These Trends in International Mathematics and Science Studies (TIMSS), conducted in 1995, 1999, and 2003, built on earlier IEA studies (Martin et al., 2004; Mullis et al., 2004), and involved over 50 countries. A significant proportion of these countries participated with the assistance of the World Bank and other development agencies. These funding agencies often wish to use the TIMSS data to monitor achievement and inform educational policy in the developing countries (Gilmore, 2005). More generally, participating countries are concerned with raising the level of student performance in their education systems; perhaps most of all in the case of their lowest performing students. This paper explores ways of summarising the performance of lower achieving students on TIMSS with a view to monitoring changes in such performance over time. The concepts and methods (e.g., the use of indices to monitor changes) used are drawn from the literature on poverty.

Sen (1976), as well as later researchers who picked up on his ideas, viewed poverty measurement as involving two steps: the identification of the poor and the aggregation of data on poverty into an overall index. By definition, a poor person was someone who fell below a poverty line, usually defined as an income level. The aggregation step involved the application of a rule or formula. The resulting index should be sensitive to inequality among the poor (Sen, 1976). One such group of indices was developed by Foster, Greer, and Thorbecke (1984), and is now widely known as the FGT indices. With a slight change in the basic index formulation, these indices can be easily adapted to describe mathematics performance above or below a particular cut-point on a test. The result is a new class of indices that are useful in monitoring changes in the performance of lower achieving students over time. The rest of this paper describes this new class of achievement indices, and then applies them to data from the TIMSS 1995, 1999, and 2003 mathematics assessments.

ADAPTING FOSTER, GREER, AND THORBECKE'S POVERTY INDICES

A competency cut -point is an achievement level such that students whose achievement is lower than the cut-point fail, and students whose achievement is equal to or higher than the cut-point pass. The difference between the failing student's score (θ_i) and the cut-point (z) can be defined as the score shortfall or deficit (g_i). If the student's score is equal to or above the cut-point, then the shortfall is zero by definition. The split function describing the computation of the shortfall score is:

$$g_i = \begin{matrix} 0 & \theta_i \geq z \\ z - \theta_i & \theta_i < z \end{matrix} \qquad (1)$$

Following Foster, Greer, and Thorbecke (1984), a number of failure indices can be developed to summarise the shortfall within any population or sub population. These indices can be represented by the following formula:

$$P_\alpha = \frac{1}{n} \sum_{i=1}^{q} \left(\frac{g_i}{z} \right)^\alpha \qquad (2)$$

Where q is the number of students inside the shortfall region and n is the sample size. The parameter α measures the sensitivity of the index to the degree of failure of those classified by the benchmark as having a value less than z, and usually assumes values of 0, 1, 2, and so on. This index nests several special cases. If $\alpha = 0$ the index is the proportion of students below the cut-point. If $\alpha = 1$, the index is the average of the proportionate shortfall gaps. When $\alpha = 2$, the proportionate shortfall gaps are weighted so that a doubling of the proportionate shortfall gap contributes four times as much to the index. And when $\alpha = 3$, a doubling of the proportionate shortfall gap contributes nine times as much to the index. Practically speaking then, a low index value when $\alpha = 0$ means that relatively few students are below the cut-point, while high index values, when $\alpha = 2$ or $\alpha = 3$, indicates that there are a significant number of students who have very low scores at some distance from the cut-point.

The Foster, Greer, and Thorbecke indices enable the specification of different poverty lines, consistent with the fact that such lines vary from country to country. However, if the same cut-point is used across countries, then the denominator can be removed from the indices with no loss of information. A further refinement lies in the sample divisor. In its current form, the indices are summed over q points, the number of students equal below the line, and then expressed in numerical terms with reference to the sample. That is, the indices are divided by the total sample size. One interpretation difficulty with this method is that if a sizeable proportion of the sample is at or above the cut point, the indices become relatively insensitive to changes below the cut-point. Another interpretation difficulty lies in the scale properties of the indices. The scale metric is lost when the indices are computed by dividing the shortfall by the cut-point. For these reasons, the following group of indices, called the modified FGT indices or P_β , is developed:

$$P_\beta = \frac{\sqrt[\beta]{\sum_{i=1}^{q} (g_i)^\beta}}{q}, \beta \geq 1 \qquad (3)$$

In this group of indices β is an integer greater than zero. When $\beta = 1$, the index is the average distance from the cut-point for those below that point. For $\beta = 2$, the index is the average of the square root of the sum of squared shortfalls and is computationally similar to the standard

deviation. When a modified index is combined with the original FGT index with $\alpha = 0$, the resulting index is expressed over the sample.

$$P = \frac{\sqrt[\beta]{\sum_{i=1}^{q} (g_i)^{\beta}}}{q} x \frac{\sum_{i=1}^{q} (g_i)^{0}}{n} = \frac{\sqrt[\beta]{\sum_{i=1}^{q} (g_i)^{\beta}}}{n}$$

or

$$P = \frac{\sqrt[\beta]{\sum_{i=1}^{q} (g_i)^{\beta}}}{q} x \frac{q}{n} = \frac{\sqrt[\beta]{\sum_{i=1}^{q} (g_i)^{\beta}}}{n} \quad (4)$$

The FGT shortfall indices are all additively decomposable. That is, each index can be decomposed to yield index values for mutually exclusive and exhaustive sub groups. For example, the indices can be decomposed to yield values for male and female students:

$$P_\alpha = P_{g\alpha} + P_{b\alpha}$$

In this manner, comparisons can be made of various sub groups of interest to policy makers and the like. However this property does not generally apply to the modified indices except in the special case when the sub-groups are of equal size and $\beta = 1$.

TIMSS MATHEMATICS ACHIEVEMENT DATA AND SHORTFALL INDICES

TIMSS used Bayesian population estimates that employ plausible or imputed values methods to overcome problems associated with distributing a large number of test items across several test booklets. The procedures used to obtain these Bayesian estimates for TIMSS 1995 and 1999 were described by Yamamoto and Kulick (2000) and Gonzalez, Galia, and Li (2004) for TIMSS 2003. The Bayesian population estimates were obtained by randomly drawing values from a distribution of possible values formed for each student. For both mathematics and science, five plausible were drawn for each assessed student. When calculated over all participating countries, the average of the five plausible values for mathematics would be 500 scale points, and the standard deviation would be 100 (using the original TIMSS scale). These mean and standard deviation statistics were calculated by computing the mean and standard deviation for each plausible value, and then calculating the average of these values

The shortfall indices, adapted to use all five plausible values for each student, are as follows:

$$g_{ik} = \begin{cases} 0 & \theta_{ik} \geq z \\ z - pv_{ik} & \theta_{ik} < z \end{cases} \quad (5)$$

where pv_{ik} is the kth plausible value for the ith student. The FGT-type indices are calculated by averaging over all plausible values.

$$P_\alpha = \frac{1}{5} \sum_{k=1}^{5} \frac{1}{n} \sum_{i=1}^{q_k} \left(\frac{g_{ik}}{z} \right)^{\alpha} \quad (6)$$

Note that the number of students falling below the cut-point can vary from plausible value to plausible value. Similar changes can be made to the modified indices to utilise the five plausible values.

Countries participating in TIMSS typically used stratified, cluster-sampling strategies (Foy, 2000). These sampling designs were considered efficient ways of obtaining representative achievement data from education systems. Typically, countries sampled intact mathematics classrooms from randomly sampled schools that were selected using a probability proportional to size method. Thus, the calculation of the indices required the use of an appropriate set of weights. In addition, the design effects associated with such sampling plans should be taken into account when calculating the standard errors of the shortfall indices. The analyses reported here use student weights and an implementation of the jackknife procedure (Gonzalez and Miles, 2001).

Cut-points are typically determined by specific educational, psychometric, or policy criteria. However, for illustrative purposes an arbitrary cut-point is chosen in this article. Since the TIMSS scales were designed to have a mean of 500, based upon a 1995 cohort, the choice of 500 as the cut-point is reasonable. This value has served as the mathematics scale reference point in the last two TIMSS assessments and was the average mathematics performance of grade 8 students participating in the 1995 assessment (Mullis et al, 2004). The analyses reported here involved the calculation of mathematics shortfall indices using $\alpha = 0$ for the FGT index and $\beta = 1$ and 2 for the modified FGT indices for those countries that participated in TIMSS 1999 and at least one of the other TIMSS assessments. In order to both simply the indices and communicate more succinctly the characteristic of each index, the following nomenclature is used:

$B_{500\ \alpha\ 0}$ - the index is referring to students below the 500 cut-point and using an alpha coefficient of zero and the original FGT formula,

$B_{500\ \beta\ 1}$ - the index is referring to students below the 500 cut-point and using the modified index with a beta coefficient of one,

$B_{500\ \beta\ 2}$ - the index is referring to students below the 500 cut-point and using the modified index with a beta coefficient of two.

Significance testing was performed using a two-tailed alpha level of 0.05, adjusted for multiple comparisons using the Bonferroni method. This was a conservative method and might serve to mask important changes at the country level.

RESULTS

When the FGT shortfall index exponent is zero, the index $\beta_{500\alpha\ 0}$ yields the percent of students whose achievement is below the 500 cut-point. As shown in Table 1, the index is fairly stable in some countries. For example, variations across the assessments of less than four percent are observed in England, Hungary, Republic of Korea, the Philippines, Romania, and the United States. In some countries, there is a sharp increase in the index from TIMSS 1995 to TIMSS 1999. In at least two of these cases, Israel and Italy, this increase can be explained by a change in the sampling coverage. In the case of Israel, the 1999 sample included Arab-speaking schools while the 1995 study did not. Interestingly, the percent of students in the shortfall region in Israel decreased from 1999 to 2003. For Italy, the 1999 sample represented the entire country while the 1995 sample represented only those provinces that chose to participate. Other countries with significant increases in students falling within the region from 1995 to 1999 included the Czech Republic, Iran, Singapore, and Thailand. Tunisia and Belgium (Flemish) showed significant increases from 1995 to 2003.

Table 1: Percent of students below the International Mathematics Mean (500) in TIMSS 1995, 1999, and 2003 ($\beta_{500\alpha\ 0}$)

Country	**1995**		**1999**		**2003**	
Australia	38.31	(1.84)	35.79	(2.55)	47.67	(2.46)
Belgium (Flemish)	23.68	(2.87)	20.12	(1.39)	26.57	(1.34)
Bulgaria	40.55	(2.51)	43.91	(2.78)	60.37	(2.07)
Canada	37.25	(1.09)	33.1	(1.01)	----	----
Chile	----	----	89.46	(1.65)	90.35	(.82)
Chinese Taipei	----	----	19.05	(1.10)	20.72	(1.4)
Cyprus	59.93	(1.07)	58.61	(0.93)	66.63	(.78)
Czech Rep.	28.77	(1.81)	40.95	(2.46)	----	----
England	50.08	(1.50)	52.54	(2.32)	52.74	(2.95)
Finland	----	----	36.24	(1.61)	----	----
Hong Kong	16.87	(2.49)	13.04	(1.63)	11.85	(1.42)
Hungary	36.38	(1.68)	34.21	(1.67)	35.21	(1.74)
Indonesia	----	----	83.21	(1.25)	84.06	(1.31)
Iran, Islamic Rep.	84.85	(1.22)	82.38	(1.34)	87.93	(0.74)
Israel	37.68	(2.95)	61.21	(1.72)	50.65	(1.69)
Italy	51.92	(1.75)	57.6	(1.83)	57.23	(1.65)
Japan	14.85	(0.54)	16.14	(.59)	17.85	(0.73)
Jordan	----	----	74.9	(1.37)	79.73	(1.50)
Korea, Rep. of	16.07	(0.72)	13.66	(0.60)	13.88	(0.59)
Latvia	54.65	(1.72)	47.41	(1.74)	44.71	(1.69)
Lithuania	61.74	(2.10)	59.32	(2.17)	48.13	(1.48)
Macedonia, Rep. of	----	----	71.62	(1.54)	75.78	(1.54)
Malaysia	----	----	41.01	(2.44)	46.4	(2.35)
Moldova, Rep. of	----	----	64.34	(1.93)	66.47	(2.00)
Morocco	----	----	97.54	(0.27)	95.53	(0.46)
Netherlands	33.48	(3.30)	25.88	(3.74)	30.36	(2.18)
New Zealand	48.64	(2.31)	52.41	(2.60)	53.35	(2.68)
Philippines	----		94.87	(0.99)	90.93	(1.30)
Romania	58.19	(2.19)	59.9	(2.46)	59.66	(2.10)
Russian Federation	36.66	(2.79)	37.66	(2.79)	45.58	(2.02)
Singapore	3.65	(0.61)	9.99	(1.60)	11.07	(1.34)
Slovak rep.	32.57	(1.56)	32.06	(2.04)	45.74	(1.75)
Slovenia	34.11	(1.51)	35.66	(1.51)	53.84	(1.27)
South Africa	94.36	(1.86)	96.16	(0.85)	95.59	(1.09)
Thailand	40.49	(2.91)	65.87	(2.49)	----	----
Tunisia	----	----	78.86	(1.21)	92.35	(0.82)
Turkey	----	----	79.39	(1.61)	----	----
United States	51.21	(2.38)	48.28	(1.81)	47.7	(1.77)

= significant increase from TIMSS 1995 = significant decrease from TIMSS 1995

= significant increase from TIMSS 1999 = significant decrease from TIMSS 1999

When the shortfall exponent is 1, the modified index $\beta_{500\beta 1}$ produces the average shortfall of those students below the cut-point. In Table 2 the results of these calculations are presented. The average shortfall ranges from a low of 26.01 (Singapore, 1995) to a high of 250.17 (South Africa, 2003). In general the average shortfall is remarkably stable across the years. For example, in 22 of the 36 countries that participated in two or more assessments**,** shown in Table 2, there is no significant change in the average shortfall. The average shortfall increased from 1995 to 1999 in Czech Republic, Israel, Singapore, and Thailand. In the case of Singapore, the average shortfall is almost doubled. The average shortfall in 2003 is higher than in 1995 in Singapore, Slovak Republic, and Slovenia. Compared with the 1999 average shortfall, the 2003 shortfall is higher in Cyprus, Slovak Republic, and Tunisia.

Downward trends in the average shortfall indicate upward trends in the achievement of students below the cut-point. Such changes are observed in Cyprus (1995 to 1999), Republic of Korea (1995 to 1999), Latvia (1995 to 2003), Lithuania (1995 to 2003), Morocco (1999 t 2003), and the Philippines (1999 to 2003). Both Morocco and the Philippines show substantial improvements in the average shortfall index.

When $\beta = 2$, the modified index $\beta_{500\beta\ 2}$ provides the average of the square root of the sum of squared shortfalls. This index is more sensitive to extreme values. Thus a number of students with very low scale scores make a disproportionately high contribution to the index compared to students closer to the cut-point. The modified shortfall index ($\beta = 2$) values are presented in Table 3. The index values range from a low of 12.68 (Singapore, 1995) to a high of 705.91 (South Africa, 2003). Significant increases in the index occur from 1995 to 1999 in Czech Republic, Singapore, Slovenia, and Thailand, while a decrease is recorded in Cyprus. Compared with the 1995 index, Singapore, Slovak Republic, and Slovenia have higher index values in 2003 while Cyprus, Italy, Latvia, and Lithuania have lower values. Tunisia and Slovak Republic have higher values in 2003 compared to 1999, while Chinese Taipei, Israel, Morocco, and the Philippines have significantly lower values. Interestingly, the Moroccan 1999 value is approximately twice than of the 2003 index, indicating a substantial improvement in the lower performing students.

The shortfall indices are particularly useful in tracking changes in performance within a population. For example, Bulgaria's mean mathematics score is seen to decline from 527 in TIMSS 1995, 511 in TIMSS 1999, to 476 in TIMSS 2003 (Mullis et al, 2004). As shown in Table 1, the percentages of students falling below 500 do not change appreciably between 1995 and 1999, but do increase markedly in 2003. From the Table 2 it is suggested that much of the change in performance from 1995 to 1999 may be attributed to a decline in performance of high performing students since there is a slight, but not significant, decrease in average shortfall in 1999 compared with 1995. However, the Bulgarian average shortfall in TIMSS 2003 is substantially larger than in the earlier assessments. From the combined data in Tables 1 and 2, it is suggested that there was a dramatic and widespread decrease in Bulgarian performance on the TIMSS 2003 mathematics assessment.

DISCUSSION

In this paper a new class of indices useful in summarising changes in achievement is presented. The new indices, based upon the Foster, Greer, and Thorbecke (1984) indices, were applied to the TIMSS mathematics data. Trends in performance below the international mean of 500 are monitored, and the new class of indices appears to be useful in detecting changes in performance over time.

Table 2: Average shortfall of students below TIMSS International Mathematics Mean for TIMSS 1995, 1999, and 2003 ($\beta_{500\beta1}$)

Country	**1995**		**1999**		**2003**	
Australia	67.09	(2.59)	59.07	(2.31)	64.84	(3.14)
Belgium (Flemish)	53.97	(6.29)	51.91	(6.62)	58.50	(3.84)
Bulgaria	67.39	(2.37)	65.99	(2.20)	77.21	(2.39)
Canada	53.59	(1.77)	50.94	(1.31)	----	----
Chile	----	----	124.92	(2.36)	129.62	(2.62)
Chinese Taipei	----	----	73.74	(2.19)	62.73	(2.15)
Cyprus	92.19	(1.67)	76.37	(1.36)	83.96	(1.39)
Czech Rep.	43.52	(1.80)	54.12	(1.75)	----	----
England	69.23	(2.11)	66.52	(2.20)	61.64	(3.25)
Finland	----	----	47.84	(1.63)	----	----
Hong Kong	63.75	(6.16)	47.31	(5.82)	49.53	(6.02)
Hungary	57.39	(2.33)	61.20	(2.12)	54.97	(2.21)
Indonesia	----	----	127.91	(3.58)	115.17	(4.05)
Iran, Islamic Rep.	102.79	(3.27)	103.43	(1.94)	105.89	(1.95)
Israel	67.48	(4.39)	91.86	(3.35)	71.36	(2.25)
Italy	79.91	(2.78)	78.47	(2.46)	68.65	(2.07)
Japan	45.52	(1.34)	48.06	(1.64)	48.20	(1.52)
Jordan	----	----	114.83	(2.12)	107.24	(2.48)
Korea, Rep. of	57.56	(2.46)	47.15	(1.39)	53.44	(1.66)
Latvia	69.01	(2.55)	60.16	(1.83)	56.81	(1.78)
Lithuania	78.16	(2.56)	69.65	(2.76)	64.52	(1.57)
Macedonia, Rep. of	----	----	97.08	(3.06)	100.22	(2.70)
Malaysia	----	----	58.23	(2.15)	56.82	(1.88)
Moldova, Rep. of	----	----	80.77	(2.03)	83.13	(2.58)
Morocco	----	----	168.66	(1.58)	119.30	(1.95)
Netherlands	56.48	(7.20)	53.53	(5.68)	46.53	(3.50)
New Zealand	67.43	(2.63)	76.69	(2.37)	65.66	(3.58)
Philippines	----	----	166.07	(4.38)	138.31	(3.88)
Romania	88.52	(2.90)	86.32	(3.60)	84.13	(2.85)
Russian Federation	59.83	(2.31)	60.20	(3.00)	58.91	(1.82)
Singapore	26.01	(1.45)	45.89	(4.06)	45.78	(2.59)
Slovak rep.	52.03	(1.66)	49.80	(1.76)	63.76	(2.16)
Slovenia	47.98	(1.41)	56.70	(2.07)	59.12	(1.42)
South Africa	238.59	(7.07)	236.28	(4.10)	250.17	(3.67)
Thailand	58.07	(2.22)	80.06	(2.34)	----	----
Tunisia	----	----	74.75	(1.30)	99.75	(1.52)
Turkey	----	----	102.17	(2.20)	----	----
United States	73.20	(2.97)	71.17	(2.01)	63.21	(1.87)

= significant increase from TIMSS 1995 = significant decrease from TIMSS 1995

= significant increase from TIMSS 1999 = significant decrease from TIMSS 1999

Table 3: Average of the square root of squared shortfalls for students below TIMSS International Mathematics Mean for TIMSS 1995, 1999, and 2003 ($\beta_{500}\beta_{2}$)

Country	**1995**		**1999**		**2003**	
Australia	74.54	(5.31)	56.90	(3.97)	66.88	(6.49)
Belgium (Flemish)	53.01	(12.20)	47.04	(14.00)	59.32	(7.69)
Bulgaria	72.10	(4.29)	69.91	(3.86)	91.85	(5.30)
Canada	47.86	(3.01)	43.02	(1.99)	----	----
Chile	----	----	205.75	(6.58)	214.72	(7.41)
Chinese Taipei	----	----	92.21	(5.49)	64.09	(4.12)
Cyprus	132.61	(4.61)	91.34	(2.88)	106.11	(3.35)
Czech Rep.	31.07	(2.49)	48.02	(2.93)	----	----
England	76.48	(4.25)	71.11	(4.39)	58.14	(5.26)
Finland	----	----	39.37	(2.78)	----	----
Hong Kong	71.66	(12.78)	42.62	(12.45)	43.92	(9.71)
Hungary	54.01	(4.23)	63.02	(4.59)	50.39	(4.43)
Indonesia	----	----	225.55	(10.48)	182.27	(11.84)
Iran, Islamic Rep.	146.44	(8.94)	149.97	(4.77)	147.92	(4.74)
Israel	82.78	(10.79)	131.08	(8.84)	79.19	(4.55)
Italy	104.85	(6.45)	96.34	(5.56)	73.09	(4.24)
Japan	36.25	(2.26)	41.30	(2.95)	39.99	(2.42)
Jordan	----	----	190.82	(6.37)	163.04	(6.61)
Korea, Rep. of	58.19	(5.15)	39.28	(2.63)	49.08	(2.99)
Latvia	73.76	(5.51)	58.23	(3.71)	50.71	(2.80)
Lithuania	94.62	(5.67)	73.73	(5.45)	64.48	(3.02)
Macedonia, Rep. of	----	----	142.65	(7.62)	147.56	(7.45)
Malaysia	----	----	55.26	(3.98)	48.86	(2.97)
Moldova, Rep. of	----	----	97.05	(4.6)	104.18	(5.73)
Morocco	----	----	358.53	(6.03)	181.09	(5.11)
Netherlands	58.13	(14.96)	49.15	(9.57)	35.91	(5.35)
New Zealand	72.74	(5.19)	90.69	(4.89)	66.51	(7.47)
Philippines	----	----	352.76	(14.84)	245.11	(11.19)
Romania	122.21	(7.33)	117.06	(8.56)	108.3	(6.40)
Russian Federation	58.19	(3.90)	61.54	(5.53)	55.3	(3.11)
Singapore	12.68	(1.38)	35.98	(5.60)	33.74	(3.29)
Slovak rep.	46.34	(2.99)	42.26	(2.70)	65.53	(4.34)
Slovenia	36.59	(1.94)	54.22	(3.61)	54.4	(2.58)
South Africa	644.64	(27.77)	651.63	(16.43)	705.91	(15.36)
Thailand	53.85	(3.87)	96.89	(4.93)	----	----
Tunisia	----	----	80.02	(2.34)	125.61	(3.28)
Turkey	----	----	145.98	(5.39)	----	----
United States	86.76	(6.34)	79.02	(3.62)	62.36	(3.47)

= significant increase from TIMSS 1995 = significant decrease from TIMSS 1995

= significant increase from TIMSS 1999 = significant decrease from TIMSS 1999

It is relatively easy to modify both classes of indices to monitor high performance. For example, if the desire were to track changes in performance above 600 scale points, then the split function would be:

$$g_{ik} = \begin{cases} pv_{ik} - 600 & \sigma_{ik} > 600 \\ 0 & \theta_{ik} \leq 600 \end{cases} . \qquad (7)$$

Reasons for changes in index cut-point values are best provided at the local level. For example, a country may wish to monitor proficiency changes in the advanced benchmarking region of a national assessment. Nevertheless, given that the TIMSS assessments are psychometrically sound, the indices used in this article appear to be useful for monitoring changes in low performance over time. The FGT index ($\alpha = 0$) captures the percentage of students within the designated region while the modified indices provide useful summarisations of the achievement data within the region.

Issues of multidimensionality of failure arise because individuals, educators, and policy makers often need to describe achievement on several individual attributes, including knowledge, problem solving, and literacy. Multidimensional failure indices can be developed that take into account the different facets of achievement. For example, the TIMSS mathematics curriculum and assessment frameworks (Robitaille et al, 1993; Mullis et al., 2003) include a number of content areas and processes. A multidimensional mathematics failure index can include dimensions for each content and process area, and can be extended to include opportunity to learn and other factors that are shown to be related to mathematics achievement. Such an approach minimises the temptation to place undue emphasis upon an overall achievement score, and yields a richer understanding likely to inform better and more direct policy decisions. The results presented in this paper can be easily conceptualised as being weighted indices of multidimensional component indices.

REFERENCES

Foster, J., Greer, J. and Thorbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, 52(3), 761-766.

Foy, P. (2000). Sampling weights. In M.O. Martin, K.D. Gregory and S.E. Stemler (Eds.), *TIMSS 1999 Technical Report* (pp. 189-202). Boston, MA: International Study Center, Boston College.

Gonzalez, E.J. and Miles, J.A. (2001). *TIMSS 1999 User Guide for the International Database*. Boston, MA: International Study Center, Boston College.

Gonzalez, E.J., Galia, J. and Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In. M.O., I.V.S. Mullis and S.J. Chrostowski, (Eds.), *TIMSS 2003 Technical Report*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low-and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). [Online] http://www.iea.nl/iea/hq/fileadmin/user_upload/WB-report.pdf [June 26, 2005].

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J. and Chrostowski, S.J. (2004). *TIMSS 2003 International Science Report Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. and Chrostowski, S.J. (2004). *TIMSS 2003 International Mathematics Report Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowski, S.J. and O'Connor, K.M. (2003). *TIMSS Assessment Framework and Specifications.* (Second Ed). Chestnut Hill, MA.: Boston College..

Robitaille, D.F., McKnight, C.C., Schmidt, W.H., Britton, E., Raizen, S. and Nicol, C. (1993). *TIMSS Monograph No. 1: Curriculum Frameworks for Mathematics and Science*. Vancouver: Pacific Educational Press.

Sen, A. (1976). Poverty: An ordinal approach to measurement. *Econometrica*, 44(2), 219-231.

Yamamoto, K. and Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, K.D. Gregory, and S.E. Stemler (Eds.), *TIMSS 1999 Technical Report* (pp. 237-263). Boston, MA: International Study Center, Boston College.

International Education Journal, 2013, Volume 8, 160-173.
ISSN 1443-1475 © 2013.

Suppressor variables and multilevel mixture modelling

I Gusti Ngurah Darmawan
School of Education, University of Adelaide *igusti.darmawan@adelaide.edu.au*

John P. Keeves
School of Education, Flinders University *john.keeves@flinders.edu.au*

A major issue in educational research involves taking into consideration the multilevel nature of the data. Since the late 1980s, attempts have been made to model social science data that conform to a nested structure. Among other models, two-level structural equation modelling or two-level path modelling and hierarchical linear modelling are two of the techniques that are commonly employed in analysing multilevel data. Despite their advantages, the two-level path models do not include the estimation of cross-level interaction effects and hierarchical linear models are not designed to take into consideration the indirect effects. In addition, hierarchical linear models might also suffer from multicollinearity that exists among the predictor variables. This paper seeks to investigate other possible models, namely the use of latent constructs, indirect paths, random slopes and random intercepts in a hierarchical model.

Multilevel data analysis, suppressor variables, multilevel mixture modelling, hierarchical linear modelling, two-level path modelling

INTRODUCTION

In social and behavioural science research, data structures are commonly hierarchical in nature, where there are variables describing individuals at one level of observation and groups or social organisations at one or more higher levels of observation. In educational research, for example, it is interesting to examine the effects of characteristics of the school, the teacher, and the teaching as well as student characteristics on the learning or development of individual students. However, students are nested within classrooms and classrooms are nested within schools, so the data structure is inevitably hierarchical or nested.

Hierarchical data structures are exceedingly difficult to analyse properly and as yet there does not exist a fully developed method for how to analyse such data with structural equation modelling techniques (Hox, 1994, as cited in Gustafsson and Stahl, 1999). Furthermore, Gustafsson and Stahl (1999) mentioned that there are also problems in the identification of appropriate models for combining data to form meaningful and consistent composite measures for the variables under consideration.

Two commonly used approaches in modelling multilevel data are two-level structural equation modelling or two-level path modelling and hierarchical linear modelling. Despite their advantages, the two -level path models currently employed do not include the estimation of cross-level interaction effects; and hierarchical linear models are not designed to take into consideration the latent constructs as well as the indirect paths. In addition, some other problems are associated with the use of HLM, such as fixed X-variables with no errors of

measurement, limited modelling possibilities and like any regression the analysis also suffers from the multicollinearity that exists among the predictor variables. The multicollinearity issue is considered in the following section because discussion of this issue is not only highly relevant, but is also rarely undertaken.

MULTICOLLINEARITY AND SUPPRESSOR VARIABLE

Since Horst (1941) introduced the concept of the 'suppressor variable', this problem has received only passing attention in the now nearly two-thirds of a century since it was first raised. In its classical rendering Conger (1974) argued that a suppressor variable was a predictor variable, that had a zero (or close to zero) correlation with the criterion, but nevertheless contributed to the predictive validity of a test.

Three types of suppressor variables have been identified. Conger (1974) labelled them as traditional, negative and reciprocal. Cohen and Cohen (1975) named the same categories classical, net, and cooperative. To describe these three types of suppression, suppose that there are the criterion variable *Y* and two predictor variables, X_1 and X_2.

Classical Suppression

A classical suppression occurs when a predictor variable has a zero correlation with the criterion but is highly correlated with another predictor in the regression equation. In other words, $r_{Y1} \neq 0$, $r_{Y2} = 0$, and $r_{12} \neq 0$. In order to understand the meaning of these coefficients it is useful to consider the Venn diagram shown in Figure 1.

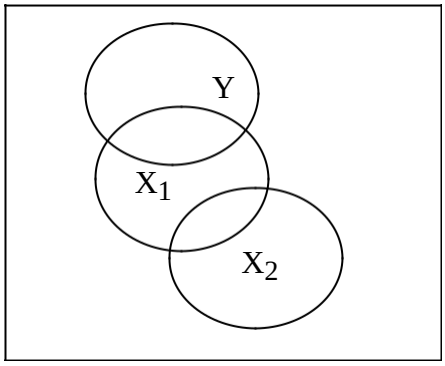


Figure 1. A Venn diagram for classical suppression

Here the presence of X_2 increases the multiple correlation (R^2), even though it is not correlated with *Y*. What happens is that X_2 suppresses some of what would otherwise be error variance in X_1.

Cohen et al. (2003, p.70) gave the formula for the multiple correlation coefficient for two predictors and one criterion as a function of their correlation coefficients:

$$R^2_{Y.12} = \frac{r_{Y1}^2 + r_{Y2}^2 - 2r_{Y1}r_{Y2}r_{12}}{1 - r_{12}^2} \qquad (3)$$

Since $r_{Y2} = 0$, equation (3) can be simplified as

$$R^2_{Y.12} = \frac{r_{Y1}^2}{1 - r_{12}^2} \qquad (4)$$

Because r^2_{12} must be greater than 0, the denominator is less than 1.0. That means that $R^2_{Y.12}$ must be greater than $r^2_{Y.1}$. In other words, even though X_2 is not correlated with Y, having it in the equation raises the R^2 from what it would have been with just X_1. The general idea is that there is some kind of noise (error) in X_1 that is not correlated with Y, but is correlated with X_2. By including X_2 this noise is suppressed (accounted for) leaving X_1 as an improved predictor of Y. The magnitude of the $R^2_{Y.12}$ depends of the values of r_{12} and $r_{1\,Y}$ as can be seen in Figure 2, where the multiple correlation ($R^2_{Y.12}$) for different values of r_{12} and for the different correlations between X_1 and Y have been presented. In some cases, the $R^2_{Y.12}$ value can be greater than 1.

Cohen et al. (2003, p. 68) gave the formula for the $\beta_{Y1.2}$ and $\beta_{Y2.1}$ coefficients as follows:

$$\beta_{Y1.2} = \frac{r_{Y1} - r_{Y2}\, r_{12}}{1 - r^2_{12}}$$

$$\beta_{Y2.1} = \frac{r_{Y2} - r_{Y1}\, r_{12}}{1 - r^2_{12}} \qquad (5)$$

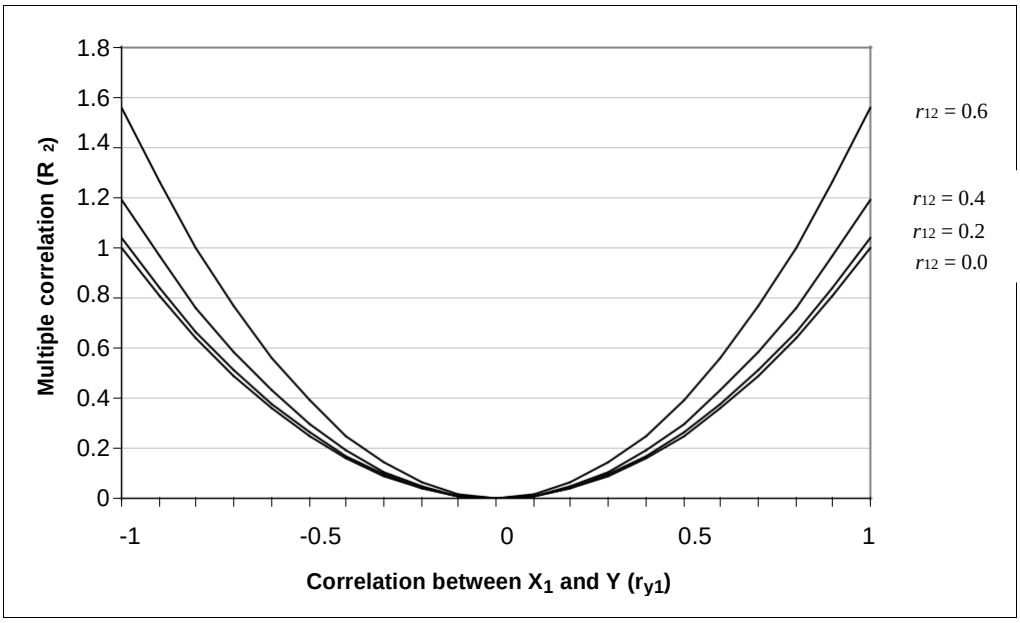


Figure 2. The inflation of $R^2_{Y.12}$

Since $r_{Y2} = 0$, Equation (5) can be simplified as

$$\beta_{Y1.2} = \frac{r_{Y1}}{1 - r^2_{12}} \quad \text{and} \quad \beta_{Y2.1} = \frac{-r_{Y1}\, r_{12}}{1 - r^2_{12}} \qquad (6)$$

The sign of $\beta_{Y2.1}$ depends on the sign of r_{12}. If there is a negative correlation between X_1 and X_2, the sign of $\beta_{Y2.1}$ will be the same as the sign of $\beta_{Y1.2}$. If there is a positive correlation between X_1 and X_2, the sign of $\beta_{Y2.1}$ and $\beta_{Y1.2}$ will be the opposite as can be seen in Figure 3. When $\beta_{Y2.1}$ has a positive sign, Krus and Wilkinson (1986) labelled it as 'positive classical suppression', and when $\beta_{Y2.1}$ has a negative sign they labelled it as 'negative classical suppression'. The magnitude of the inflations of $\beta_{Y2.1}$ and $\beta_{Y1.2}$ from their bivariate correlation with the criterion, r_{Y1} and r_{Y2} also depend on the value of r_{12}. A higher the value of

r_{12} leads to bigger inflations of $\beta_{Y1.2}$ and $\beta_{Y2.1}$ and beyond a certain point the value of $\beta_{Y1.2}$ and $\beta_{Y2.1}$ can exceed 1.

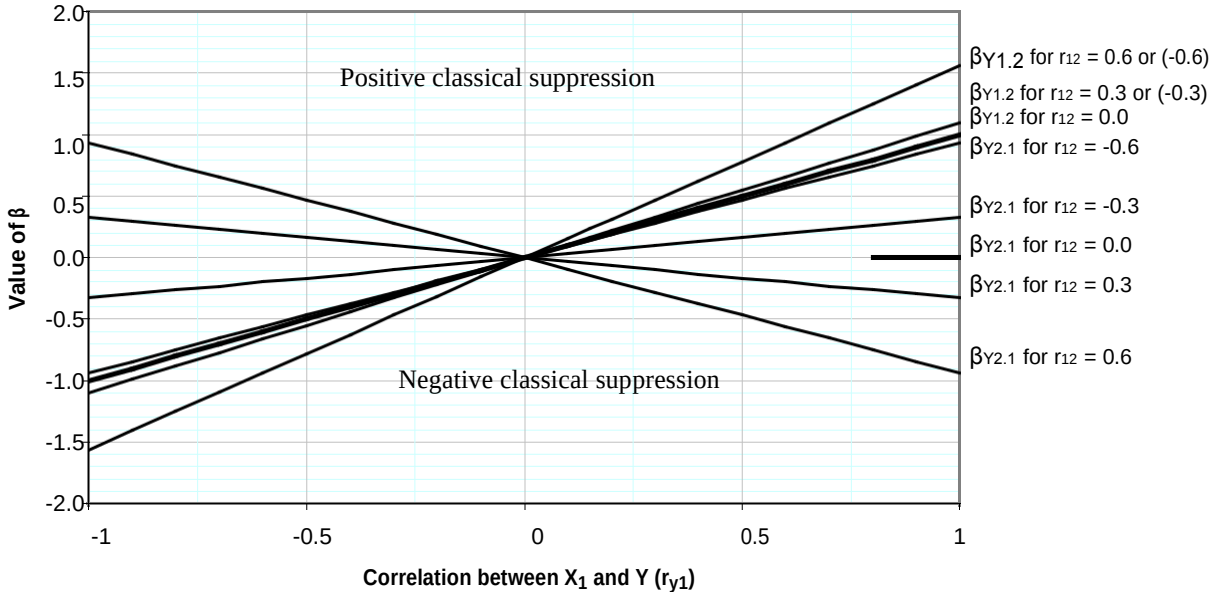


Figure 3. Classical suppression

Net suppression

This type of suppression occurs when a predictor variable has a regression weight with an opposite sign to its correlation with the criterion. In other word, $r_{Y1} \neq 0$, $r_{Y\,2} \neq 0$, and $r_{12} \neq 0$ but the $\beta_{Y2.1}$ is opposite in sign to r_{Y2}. In order to understand the meaning of these coefficients it is useful to consider the Venn diagram shown in Figure 4.

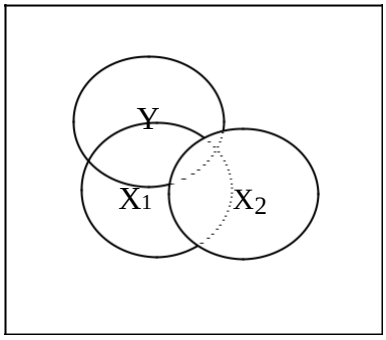


Figure 4. A Venn diagram for net suppression

Here the primary function of X_2 is to suppress the error variance X_1 , rather than influencing substantially Y. As can be seen in Figure 4 X_2 has much more in common with the error variance in X_1 than it does with the variance in Y. This can happens when X_2 is highly correlated with X_1 but weakly correlated with Y.

In Figure 5 various $\beta_{Y2.1}$ values for $r_{12} = 0.6$ and $r_{12} = -0.6$ have been plotted. If X_2 is positively correlated with Y but has a negative value of $\beta_{Y2.1}$, Krus and Wilkinson (1986) labelled it as 'negative net suppression'. If X_2 is negatively correlated with Y but has a positive value of $\beta_{Y2.1,}$ Krus and Wilkinson (1986) called it 'positive net suppression'.

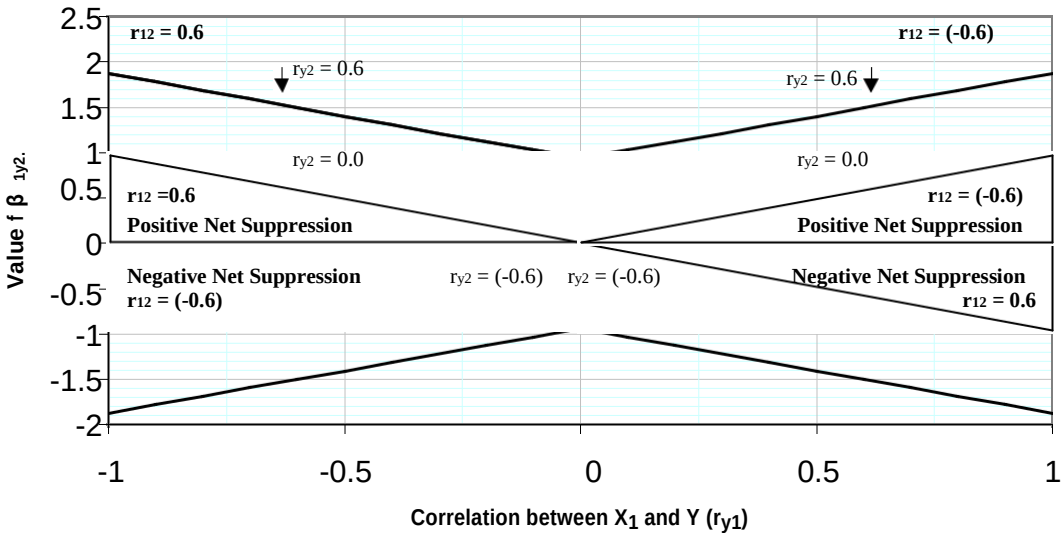


Figure 5. Net Suppression

Cooperative suppression

Co-operative suppression occurs when the two predictors are negatively correlated with each other, but both are positively or negatively correlated with *Y*. This is a case where each variable accounts for more of the variance in *Y* when it is in an equation with the other than it does when it is presented alone. As can be seen in Figure 6, when r_{12} is set to -0.6, the value of R^2 is more highly boosted as r_{Y2} increases. When both X_1 and X_2 are positively correlated with *Y*, Krus and Wilkinson (1986) labelled it as "positive cooperative suppression"; and when both X_1 and X_2 are negatively correlated with *Y*, Krus and Wilkinson (1986) labelled it as 'negative cooperative suppression' as shown in Figure 7.

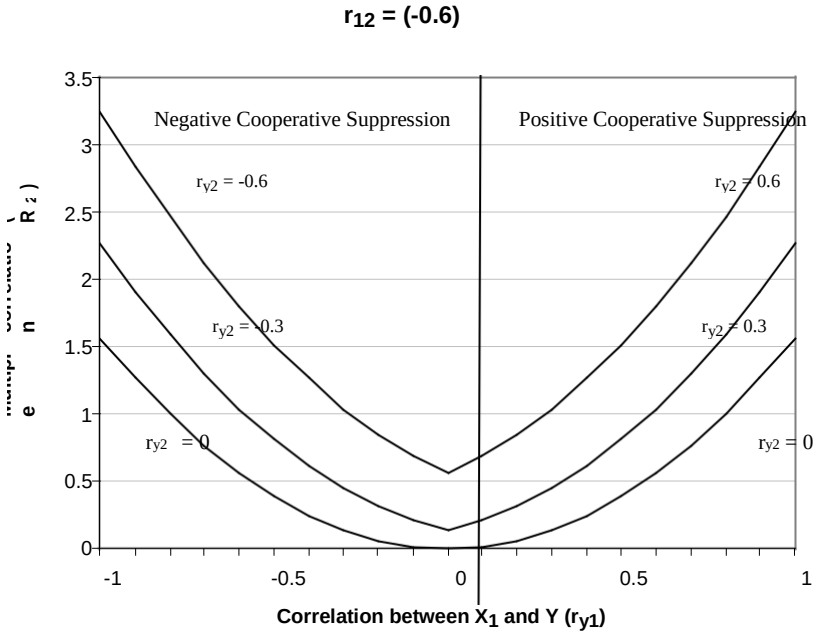


Figure 6. R^2 values in Cooperative Suppression

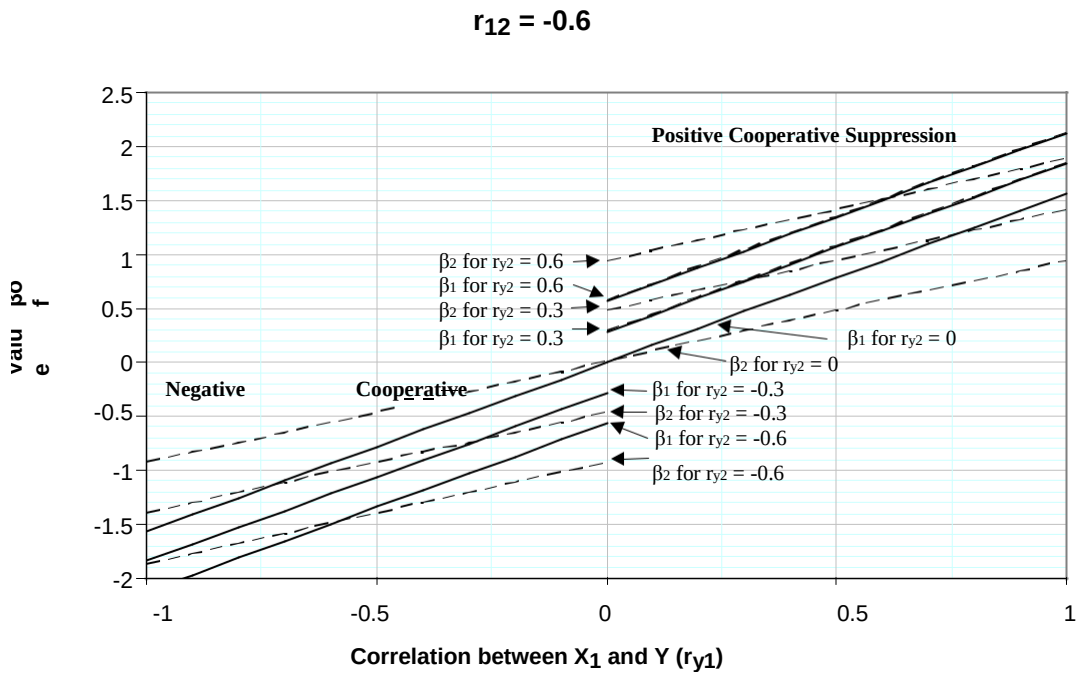


Figure 7. Cooperative Suppression

Cohen and Cohen (1983) suggested that one indication of suppression is a standardised regression coefficient (β_i) that falls outside the interval $0 < \beta_i < r_{Yi}$. To paraphrase Cohen and Cohen (1983), if X_i has a (near) zero correlation with Y, then there is possible classical suppression present. If its b_i is opposite in sign to its correlation with Y, there is net suppression present. And if its b_i exceeds r_{Yi} and it has the same sign, there is cooperative suppression present.

Multicollinearity has adverse effects not only on the regression and the multiple correlation coefficients, but also on the standard errors of regression coefficients as well as on the accuracy of computations due to rounding errors. In order to detect such problems concepts of a 'variance inflation factor' (VIF) and 'tolerance' were introduced (Pedhazur, 1997; Cohen et al., 2003).

$$VIF_i = \frac{1}{1 - R_i^2}$$
$$Tolerance = \frac{1}{VIF_i} = 1 - R^2 \qquad (7)$$

For a regression with two independent variables:

$$R^2_1 = R^2_2 = 1 - \frac{1}{\frac{1}{1 - r^2_{12}}} = 1 - (1 - r^2_{12}) = r^2_{12} \qquad (8)$$

$$VIF_1 = VIF_2 = \frac{1}{1 - r^2_{12}}$$
$$Tolerance_1 = Tolerance_2 = 1 - r_{12}^2 \qquad (9)$$

The smaller the tolerance or the higher the VIF, the greater are the problems arising from multicollinearity. There is no agreement on cut-off values of tolerance. BMDP uses a tolerance of 0.01 as a default cut-off for entering variables, MINITAB and SPSS use a default value of 0.0001 (Pedhazur, 1997, p. 299). Cohen et al. (2003, p. 423) suggested that any VIF of 10 or more provides evidence of serious multicollinearity, which is equal to a tolerance of 0.1. Furthermore, they argued that "the values of the multicollinearity indices at which the interpretation of regression coefficients may become problematic will often be considerably smaller than traditional rule of thumb guidelines such as VIF =10". Sellin (1990) used the squared multiple correlation between a predictor and the set of remaining predictors involved in the equation (R_i^2) to indicate the relative amount of multicollinearity, He mentioned that relatively large values, typically those larger than 0.5, which is equal to VIF = 2, may cause problems in the estimation.

SOME ALTERNATIVE STRATEGIES

When a researcher is concerned only with the prediction of Y, multicollinearity has little effect and no remedial action is needed (Cohen et al., 2003 p.425). However, if interest lies in the value of regression coefficients or in the notion of causation, multicollinearity may introduce a potentially serious problem. Pedhazur (1997) and Cohen et al. (2003) proposed some strategies to overcome this problem that included (a) model respecification, (b) collection of additional data, (c) using ridge regression, and (d) principal components regressions.

When two or more observed variables are highly correlated, it may be possible to create a latent variable, that can be used to represent a theoretical construct which cannot be observed directly. The latent construct is presumed to underlie those observed highly correlated variables (Byrne, 1994).

The authors of this article have focused on this strategy, to create latent constructs and to extend the hierarchical linear model to accommodate the latent constructs. It also seeks to include indirect paths into the hierarchical linear model with the latent predictor. Thus, an attempt has been made to combine the strengths of the two common approaches in analysing multilevel data: (a) two-level path models that can estimate direct and indirect effects at two levels, can use latent constructs as predictor variables, but can not estimate any cross-level interaction; and (b) hierarchical linear models that can estimate direct and cross-level interaction effects, but can not estimate indirect paths nor use latent constructs as predictor variables. Muthén and Muthén (2004) have developed a routine called 'multilevel mixture modelling' that can estimate a two-level model which has latent constructs as predictor variables, direct and indirect paths, as well as cross-level interactions.

DATA AND VARIABLES

The data used in this study were collected from 1,984 junior secondary students in 71 classes in 15 schools in Canberra, Australia. Information was collected about individual student socioeconomic status (father's occupation), student aspirations (expected occupation, educational aspirations (expected education), academic motivation, attitude towards science (like science), attitude towards school in general (like school), self-regard, prior science achievement and final science achievement (outcome). In addition, information on class sizes was also collected. The outcome measure was the scores on a science achievement test of 55 items.

The names, codes and description of the predictor variables tested for inclusion at each level have been given in Table 1.

Table 1: Variables tested at each level of the hierarchy

Level	Variable code	Variable description
Level-1		(Student-level)
Student Background (N=1984)	FOCC	Father's occupation (1=Professional, . . . , 6=Unskilled labourer)
	EXPOCC	Expected occupation (1=Professional, . . . , 6=Unskilled labourer)
	EXPED	Expected education (1=Year 10 and Below, . . . ; 6=Higher Degree)
	ACAMOT	Academic motivation (0=Lowest motivation, . . . , 40=Highest motivation)
	LIKSCH	Like school (0=Likes school least, . . . , 34=Likes school most)
	LIKSCI	Like science (1=Likes science least, . . . , 40=Likes science most)
	SELREG	Self regard (1=Lowest self regard, . . . , 34=Highest self regard)
	ACH68	Prior science achievement (0=Lowest score, . . . , 25=Highest score)
Level-2		(Class-level)
Class Characteristics Group Composition (n=71)	CSIZE	Class size (8=Smallest, . . . , 39=Largest)
	FOCC_2	Average father occupation at class-level
	EXPOCC_2	Average expected occupation at class-level
	EXPED_2	Average expected education at class-level
	ACAMOT_2	Average academic motivation at class-level
	LIKSCH_2	Average like school at class-level
	LIKSCI_2	Average like science at class-level
	SELREG_2	Average self regard at class-level
	ACH68_2	Average prior science achievement
Outcome	ACH69	Science Achievement (1 =lowest score….55=highest score)

HLM MODEL: THE INITIAL MODEL

Initially a two-level model was fitted using HLM 6. The first step in the HLM analyses was to run a fully unconditional model in order to obtain the amounts of variance available to be explained at each level of the hierarchy (Bryk and Raudenbush, 1992). The fully unconditional model contained only the dependent variable (Science achievement, ACH) and no predictor variables were specified at the class level. The fully unconditional model is stated in equation form as follows.

Level-1 model

$$Y_{ij} = \beta_{0j} + e_{ij}$$

Level-2 model

$$\beta_{0j} = \gamma_{0j} + r_{0j} \qquad (10)$$

where:

Y_{ij} is the science achievement of student *i* in class *j*;

The second step undertaken was to estimate a Level-1 model, that is, a model with student-level variables as the only predictors in Equation 10. This involved building up the student-level model or the so-called 'unconditional' model at Level-1 by adding student-level predictors to the model, but without entering predictors at the other level of the hierarchy. At this stage, a step-up approach was followed to examine which of the eight student-level variables (listed in Table 1) had a significant (at $p \leq 0.05$) influence on the outcome variable,

ACH69. Four variables (FOCC, EXPED, LIKSCI and ACH68) were found to be significant and therefore were included in the model at this stage. These four student-level variables were grand-mean-centred in the HLM analyses so that the intercept term would represent the ACH69 score for student with average characteristics.

The final step undertaken was to estimate a Level-2 model, which involved adding the Level-2 or class-level predictors into the model using the step-up strategy mentioned above. At this stage, the Level-2 exploratory analysis sub-routine available in HLM 6 was employed for examining the potentially significant Level-2 predictors in successive HLM runs. Following the step-up procedure, two class-level variables (CSIZE and ACH68_2) were included in the model for the intercept. In addition, one cross-level interaction effect between ACH68 and CSIZE was included in the model.

The final model at Levels 1, and 2 can be denoted as follows.

Level-1 Model

$$Y_{ij} = \beta_{0j} + \beta_{1j}*(FOCC) + \beta_{2j}*(EXPED) + \beta_{3j}*(LIKSCI) + \beta_{4j}*(ACH68) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(ACH68\_2) + \gamma_{02}*(CSIZE) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41}*(CSIZE) + u_{4j} \quad (11)$$

The next step was to re-estimate the final model using the MPLUS program. The results of the estimates of fixed effects from the two-level model are given in Table 2 for HLM and MPLUS estimation.

RESULTS

At the student-level, from the results in Table 2 it can be seen that Science achievement was directly influenced by Father's occupation (FOCC), Expected education (EXPED), Like science (LIKSCI) and Prior achievement (ACH68). When other factors were equal, students whose fathers had high status occupations (e.g. medical doctors and lawyers) outperformed students whose fathers had low status occupations (e.g. labourer and cleaners). Students who aspired to pursue education to high levels were estimated to achieve better when compared to students who had no such ambitions, while students who liked science were estimated to achieve better when compared to students who did not like science. In addition, students who had high prior achievement scores were estimated to achieve better than students who had low prior achievement scores.

At the class-level, from the results in Table 2 it can be seen that Science achievement was directly influenced by Average prior achievement (ACH68_2) and Class size (CSIZE). When other factors were equal, students in classes with high prior achievement scores were likely to achieve better when compared to students in classes with low prior achievement scores. Importantly, there was considerable advantage (in term of better achievement in science) associated with being in larger classes. These relationships have been shown in Figure 8.

From the results in Table 2 it can also be seen that there is one significant cross-level interaction effect ACH68 and CSIZE. This interaction is presented in Figure 9. Nevertheless, in interpreting the effects of class size, it should be noted that 10 out of the 15 schools in these data had a streaming policy that involved placing high achieving students in larger classes and low achieving students in smaller classes for effective teaching. Therefore, the better performance of the students in larger classes in these data was not surprising.

Table 2. HLM and MPLUS results for initial model

Level 1 N=1984	Level 2 n=71	HLM Estimate (se)	MPLUS Estimate (se)
Intercept		28.37 (0.20)	28.87 (0.19)
	ACH68_2	0.78 (0.10)	0.76 (0.12)
	CSIZE	0.16 (0.04)	0.16 (0.04)
FOCC		-0.25 (0.09)	-0.24 (0.10)
EXPED		0.48 (0.09)	0.49 (0.09)
LIKSCI		0.15 (0.01)	0.15 (0.01)
ACH68		0.91 (0.04)	0.93 (0.04)
	CSIZE	0.013 (0.005)	0.015 (0.006)

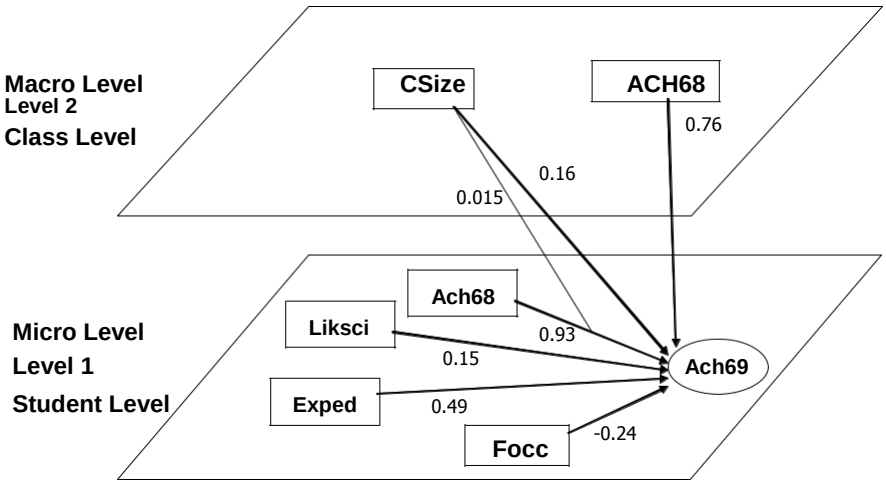


Figure 8. Model 1: Initial Model (MPlus results used)

ALTERNATIVE MODELS

Two alternative models, Model 2 and Model 3, were estimated using MPLUS 3.13. Both EXPED and EXPOCC are significantly correlated with ACH69 with correlation coefficients of 0.50 and 0.35 respectively. Either EXPED or EXOCC can have a significant effect on ACH69. However, if the two variables were put together as predictors of ACH69, only EXPED was found to be significant. Since there is a relatively high correlation between EXPOCC and EXPED (-0.53) it is possible to form a latent construct, labelled as aspiration (ASP), and use this construct as a predictor variable instead of just using either EXPOCC or EXPED. In this way, both variables (EXPOCC and EXPED) become significant reflectors of aspiration. Otherwise, EXPOCC may be regarded as an insignificant predictor of science achievement as in the initial model. The results have been recorded in Table 3 and Model 2 is shown visually in Figure 10. This employment of a latent construct is very useful in

situations where three observed predictor variables are available and suppressor relationships occur if all three predictor variables are introduced separately into the regression equation.

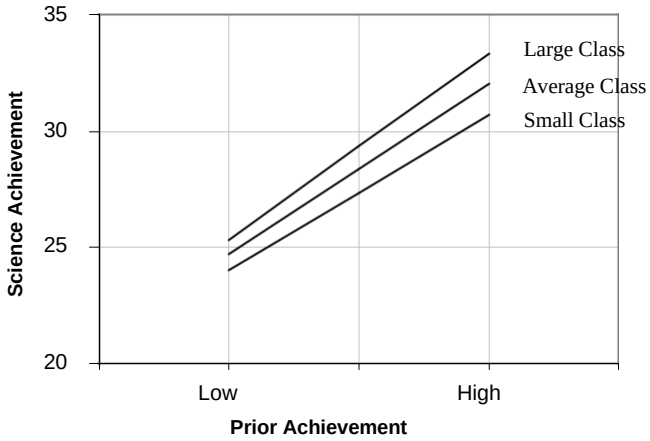


Figure 9. Interaction effect between CSIZE and PRIORACH

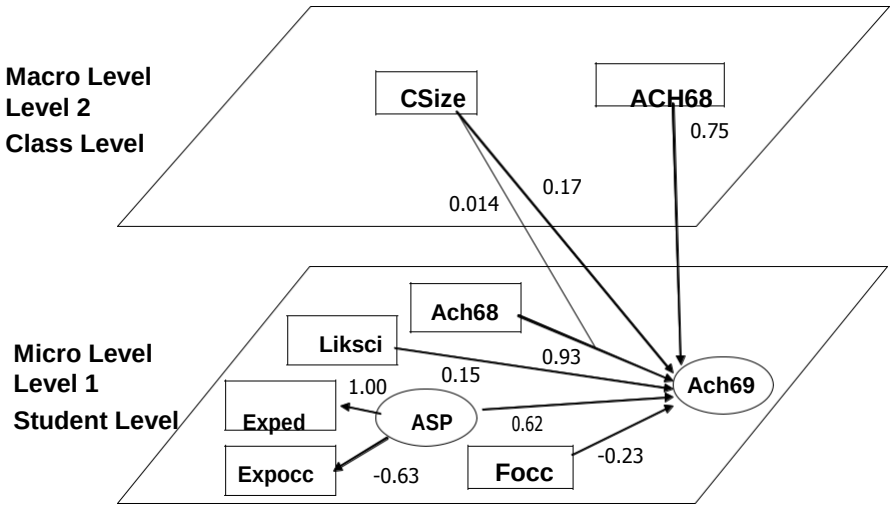


Figure 10. Model 2: With latent construct

The next step undertaken was to estimate another model with two additional indirect paths. It was hypothesised that academic motivation (ACAMOT) influenced like science at the student level and average father's occupational status influences average prior achievement at the class level. The results are recorded in Table 3 and Model 3 is shown in Figure 11.

The proportions of variance explained at each level for each model are presented in Table 4. For Model 1, the initial model, 45 per cent of variance available at Level 1 and almost all (95 %) of variance available at Level 2 have been explained by the inclusion of four variables at Level 1 (FOCC, EXPED, LIKSCI, and ACH68) and two variables at Level 2 (ACH68 and CSIZE) as well as one interaction effect between ACH68 and CSIZE. Overall this model explained 68.7 per cent of total variance available when the model was estimated with HLM. MPLUS estimations are very close to HLM estimations. Adding a latent construct into the model did not really increase the amount of variance explained, but it did give a more coherent picture of the relationships. This is also true for Model 3 when indirect paths are added.

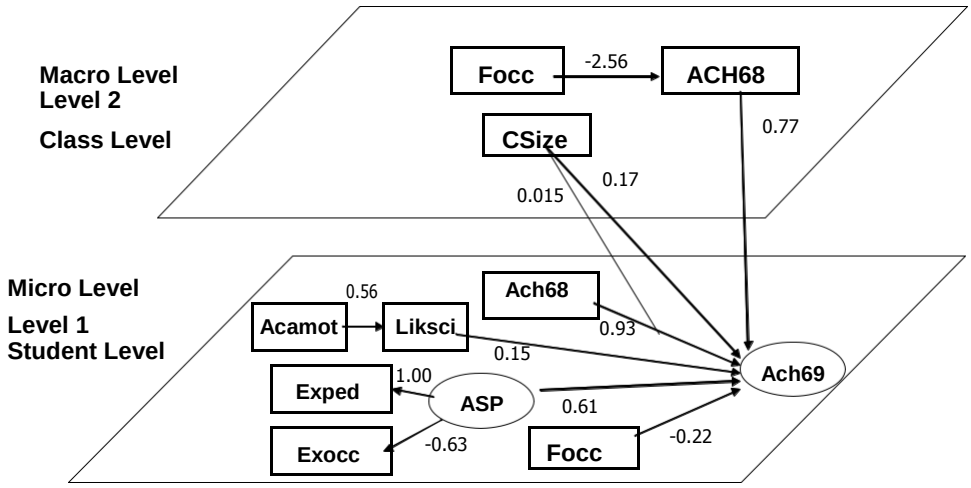


Figure 11. Model 3: With latent construct and indirect paths

Table 3. Model 2 and Model 3 Results

Level 1 (N=1984)	**Level 2 (n=71)**	**Model 2 with latent construct**	**Model 3 with latent construct and indirect paths**
Criterion ACH69		**estimate (se)**	**estimate (se)**
Latent Construct			
ASP by			
EXPED		1.00(0.00)	1.00(0.00)
EXPOCC		-0.63(0.10)	-0.63(0.11)
Indirect Paths			
ACAMOT on LIKSCI			0.56 (0.03)
	FOCC_2 on ACH68		-2.56 (0.30)
Fixed Effects			
Intercept		28.87 (0.20)	28.85 (0.20)
	ACH68	0.75 (0.12)	0.77 (0.12)
	CSIZE	0.17 (0.05)	0.17 (0.05)
FOCC		-0.23 (0.10)	-0.22 (0.10)
ASP		0.62 (0.14)	0.61 (0.14)
LIKSCI		0.15 (0.01)	0.15 (0.01)
ACH68		0.93 (0.04)	0.93 (0.04)
	CSIZE	0.014 (0.007)	0.015 (0.01)

CONCLUSIONS

Multicollinearity is one of the problems that need to be examined carefully when a multiple regression model is employed. When the main concern is merely the prediction of *Y*, multicollinearity generally has little effect, but if the main interest lies in the value of regression coefficients, multicollinearity may introduce a potentially serious problem.

Multilevel mixture modelling, which can estimate a two-level model that has latent constructs as predictor variables, direct and indirect paths, as well as cross-level interactions, has been used as an alternative strategy to analyse multilevel data. In a sense, this approach can be seen as an attempt to combine the strengths of the two commonly used techniques in analysing multilevel data, two level path modelling and hierarchical linear modelling.

The initial model was a hierarchical linear model, which was fitted using both HLM 6 and MPLUS 3.13. Both estimations yielded similar results. The main effects reported from the

analysis at the student-level, indicate that in addition to prior achievement, it was the social psychological measures associated with the differences between students within classrooms that were having effects, namely, socioeconomic status, educational aspirations, and attitudes towards learning science. About 55 per cent of the variance between students within classrooms was left unexplained, indicating that there were other student-level factors likely to be involved in influencing student achievement.

Table 4. Variance components

	HLM			**MPLUS**		
Model (N=1984, n=71)	**Level 1**	**Level 2**	**Total**	**Level 1**	**Level 2**	**Total**
Null Model						
Variance Available	38.07	33.85	71.92	38.07	33.35	71.42
Initial Achievement (Residual)	24.25	9.34	33.59	24.33	8.45	32.78
Total Variance Explained %	36.3	72.4	53.3	36.1	74.7	54.1
Total Variance Unexplained %	63.7	27.6	46.7	63.0	25.3	45.9
Model 1: Initial Model (Residual)	20.93	1.60	22.53	21.01	1.46	22.46
Total Variance Explained %	45.0	95.3	68.7	44.8	95.6	68.6
Total Variance Unexplained %	55.0	4.7	31.3	55.2	4.4	31.4
Model 2: With Latent Predictor (Residual)				21.36	1.49	22.84
Total Variance Explained %				43.9	95.5	68.0
Total Variance Unexplained %				56.1	4.5	32.0
Model 3: Add indirect Paths (Residual)				21.36	1.50	22.85
Total Variance Explained %				43.9	95.5	68.0
Total Variance Unexplained %				56.1	4.5	32.0

At the classroom level, about 4.7 per cent of the variance between classes was left unexplained, with the average level of prior achievement of the class group had a significant effect. In addition, class size had a positive effect on science achievement, with students in larger classes doing significantly better than students in smaller classes. Perhaps, this indicates the confounding effect of streaming policy adopted by some schools to place better students in larger classes. In addition, the interaction effect also reveals that the effect of prior achievement is stronger in larger classes. High achieving students are better off in larger classes.

The next step was to add a latent construct, aspiration to the initial model. The estimation of this model was done by using the two-level mixture model procedure in MPLUS 3.13. By creating this latent construct, it could be said that aspiration, which was reflected significantly by expected education and expected occupation, had a positive effect on achievement.

The last step was to add two indirect paths, one at the student level and one at the class level. At the student level, academic motivation was found to have a significant effect on like science and indirectly influence achievement through like science. At the class level, average fathers' occupation was related to average prior achievement.

By using multilevel mixture modelling, the limitations of hierarchical linear modelling are partly reduced. The ability to include latent constructs in a path model reduces the problem of multicollinearity and multiple measures. The inclusion of indirect paths also increases the modelling possibilities. However, these estimations need greater computing power if larger models are to be examined.

REFERENCE

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA.: Sage Publications.

Byrne, B.M. (1994) *Structural Equation Modelling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming*, Thousand Oaks, CA: Sage Publications.

Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: Wiley.

Cohen, J. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, (2nd ed.). Hillsdale, NJ. : Erlbaum Associates.

Cohen, J., West, S.G., Aiken, L. and Cohen, P. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed. Mahwah, NJ.: Erlbaum Associates

Conger, A.J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34, 35-46.

Gustafsson, J.E., and Stahl, P.A. (1999). *STREAMS User's Guide Version 2.5 for Windows*, Molndal, Sweden: Multivariate Ware.

Horst, P. (1941). The role of predictor variables which are independent of the criterion. *Social Science Research Bulletin*, 48, 431 -436.

Hox, J.J. (1994). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.

Krus, D.J. and Wilkinson, S.M. (1986). Demonstration of properties of a suppressor variable. *Behavior Research Methods, Instruments, and Computers*, 18, 21-24.

Muthén, L.K. and Muthén, B.O. (2004). *Mplus: The Comprehensive Modelling Program for Applied Researchers. User's guide* (3rd ed.). Los Angeles: Muthén and Muthén.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction* (3rd ed.). Forth Worth, TX.: Harcourt Brace College Publishers.

Sellin, N. (1990). *PLSPATH Version 3. 01. Application Manual*. Hamburg, Germany.

International Education Journal, 2013, Volume 8, 174-188.
ISSN 1443-1475 © 2013

Accountability of teachers and schools: A value-added approach

I Gusti Ngurah Darmawan
School of Education, University of Adelaide *igusti.darmawan@adelaide.edu.au*

John P. Keeves
School of Education, Flinders University *john.keeves@flinders.edu.au*

Currently, there has been substantial interest, in Australia and internationally, in policy activities related to outcomes-based educational performance indicators and their link with growing demands for accountability of teachers and schools. In order to achieve a fair comparison between schools, it is commonly agreed that a correction should be made for lack of equity. It is argued that student performance is influenced by three general factors: the student background, classroom and school context, and identified school policies and practices. In this article the effects of these three factors on science achievement among students in Canberra, Australia have been addressed. The effects are discussed with reference to Type A, Type B, Type X, and Type Z effects. Type A effects are school effectiveness indicators controlled for student background. Type B school effects are controlled for both student background and context variables. Type X effects are estimated with student effects, context effects and non-malleable policy effects controlled for. Finally, Type Z effects invoke school effectiveness indicators, controlled for student, context, and all identified policy effects.

Value-added, accountability, science achievement, social psychological measures, equity, school effectiveness indicators

ACCOUNTABILITY OF TEACHERS AND SCHOOLS

During the past two decades there has been a growing interest in the performance and accountability of teachers and schools both in Australia and internationally (Rowe, 2000). Educational outcome indicators are frequently used to measure the performance of teachers, schools, programs, and policies. Reliance on such indicators is largely the result of a growing demand to hold these entities accountable for performance, defined in terms of outcomes, such as standardised test scores in science, rather than inputs such as student prior achievement, teacher quality, class size, or quality of facilities (Meyer, 2000, 2002). The use of such indicators, for example average or median test scores, has some major shortcomings. Rowe (2000) pointed out that the analyses of test scores tended to be focused on a comparative ranking of schools rather than on identifying factors that explained school differences. Moreover, Meyer (2002) contended that average test scores (a) were influenced by factors other than school performance; (b) were a reflection of the accumulated learning that had occurred; (c) tended to be contaminated due to student mobility; and (d) failed to localise school performance to a specific classroom or grade level.

Given these problems associated with the use of common educational outcome indicators, the papers by Ballou et al. (2004), De Fraine et al. (2002), Raudenbush and Willms (1995), Rubin et al. (2004), and Willms and Raudenbush (1989) have approached the estimation of school and teacher effects through the use of a variety of statistical models, known as 'value-added' models in the education literature. The essence of the value-added approach is to isolate statistically the contribution of teachers and schools to growth in student achievement at a given grade level from all other sources of student achievement growth. Failure to isolate these contributions could result in highly contaminated indicators of performance.

Consequently, the emphasis in cross-national achievement surveys, as well as national studies of educational achievement that compare the performance of schools using the rank ordering or scaling of outcomes fail to examine in a meaningful way differences in performance unless further analyses that estimate value-added effects are carried out.

FOUR TYPES OF SCHOOL EFFECTS

Type A, Type B, Type X, and Type Z Effects

Raudenbush and Willms (1995, p. 313) and Willms and Raudenbush (1989, pp. 212-214) argued that student performance (Y) was influenced by three general factors: the student background characteristics (S), school context (C) and identified school policies, practices, and stratifications (P), as well as each student's unique contribution (e).

$$Y_{ij} = \mu_{0j} + S_{ij} + C_{ij} + P_{ij} + e_{ij} \quad (1)$$

This model can be extended to accommodate classroom or teacher effects by splitting school context (C) into its components, namely classroom context (CC) and school context (SC). Furthermore, school policies and practices (P) can be divided into identified policies and practices (IP) and unidentified policies and practices (UP). Identified policies and practices (IP) can be further subdivided into malleable policies and practices (MP) and non-malleable policies and practices (NP). It should be noted that non-malleable polices and practices (NP) can be identified, but a school has no control over them since they are determined at the system level, while malleable policies and practices (MP) are under a school's control. Hence we may write

$$Y_{ijk} = \mu_{0jk} + S_{ijk} + CC_{ijk} + SC_{ijk} + NP_{ijk} + MP_{ijk} + UP_{ijk} + e_{ijk} \quad (2)$$

Equation (2) can also be written with further error terms (u_{00k} and r_{0jk}) included:

$$Y_{ijk} = \gamma_{000} + S_{ijk} + CC_{ijk} + SC_{ijk} + NP_{ijk} + MP_{ijk} + UP_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \quad (3)$$

Four types of teachers or school effects can be distinguished: Type A, Type B effects (Raudenbush and Willms, 1995; Willms and Raudenbush, 1989), Type X effects (Hungi, 2003; Keeves et al., 2005) and Type Z effects.

Type A effects refer to how well the students in a school perform in comparison with the performance of similar students in other schools. Type A effects are of interest for students and parents in choosing a school. Parents want to know which school can help their child to excel. Parents and students will choose the school with the largest Type A effect, that is the school with the largest value added effect when individual student characteristics are taken into account. The Type A effects can be specified as:

$$A_{ijk} = CC_{ijk} + SC_{ijk} + NP_{ijk} + MP_{ijk} + UP_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \quad (4)$$

Type B effects refer to how well the students in a classroom within a school perform, compared to similar students in classrooms and schools with similar contexts. Type B effects are of interest for those who are looking for accountability of the teacher and school. Teachers and principals are more interested in the Type B effects of their own schools because they look for an indication of their school's performance, excluding factors that lie beyond their control. Type B effects are also of interest for administrators and education policy makers, looking for accountability. Schools should not be held accountable for the context in which they operate. The Type B effects can be specified as:

$$B_{ijk} = NP + MP_{ijk} + UP_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \quad (5)$$

There are some non-malleable polices and practices (NP), polices and practices that can be identified but the school has no control over them, and they should be removed, such as whether the school is urban or rural, or the size of the school in situations where the school has no control over its size, as well as other stratifying variables such as State or School Type. Therefore, Type X effects refer to how well the students in a classroom within a school perform, when compared to similar students in classrooms and schools with similar contexts as well as similar non-malleable policies and practices. It may be argued that the Type X estimate is the most appropriate estimate of value added, with student effects, context effects (CC and CS), and identified non-malleable policy effects (NP) removed from the value added estimates. Type X effects can be specified as:

$$X_{ijk} = MP_{ijk} + UP_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \quad (6)$$

However, it would seem appropriate to judge a school by the effect of identified malleable polices and practices as well. An example of malleable policy and practice at the school level would seem to be that of 'streaming'. After controlling for the malleable policy and practices, the remaining effects can be labelled as Type Z effects and can be written as

$$Z_{ijk} = UP_{ijk} + u_{00k} + r_{0jk} + e_{ijk} \quad (7)$$

DATA SAMPLE

The data used in this study were collected from 1,984 junior secondary students in 71 classes in 15 schools in Canberra, Australia. These 15 schools consisted of nine government schools, four Catholic schools and two independent schools. Nine of these schools were co-educational schools and six were single sex (three boys' and three girls' schools). In addition, ten out of the 15 schools had a streaming policy of placing high achieving students in larger classes. The sample represents a cohort of approximately 2000 students, who transferred from Grade 6 to Grade 7 within a small school system.

HYPOTHESIZED MODEL

Testing of hypotheses in multilevel models can be carried out using multilevel data analyses software such as HLM 6 for Windows (Raudenbush et al., 2004). The HLM program was initially developed to find a solution for the methodological weakness of educational research studies during the early 1980s, which was the failure of many analytical studies to attend to the hierarchical, multilevel character of much of educational field research data (Bryk and Raudenbush, 1992). This failure came from the fact that "the traditional linear models used by most researchers require the assumption that subjects respond independently to educational programs" (Raudenbush and Bryk; 1994, p. 2590). In practice, most educational research studies select students as a sample who are nested within classrooms, and the classrooms are in turn nested within schools, and schools exist within geographical regions. In this situation, the students

selected in the study are not independent, but rather nested within organisational units and ignoring this fact results in the problems of "aggregation bias and misestimated precision" (Raudenbush and Bryk, 1994, p. 2590).

In Figure 1 the three-level model proposed for testing in this study is shown. The names, codes and description of the predictor variables tested for inclusion at each level of the three-level model have been provided in Table 1. Apart from Class size (CSIZE) at class level and school classifications at school level, all the other variables at the class and school levels were constructed by aggregating the student-level data.

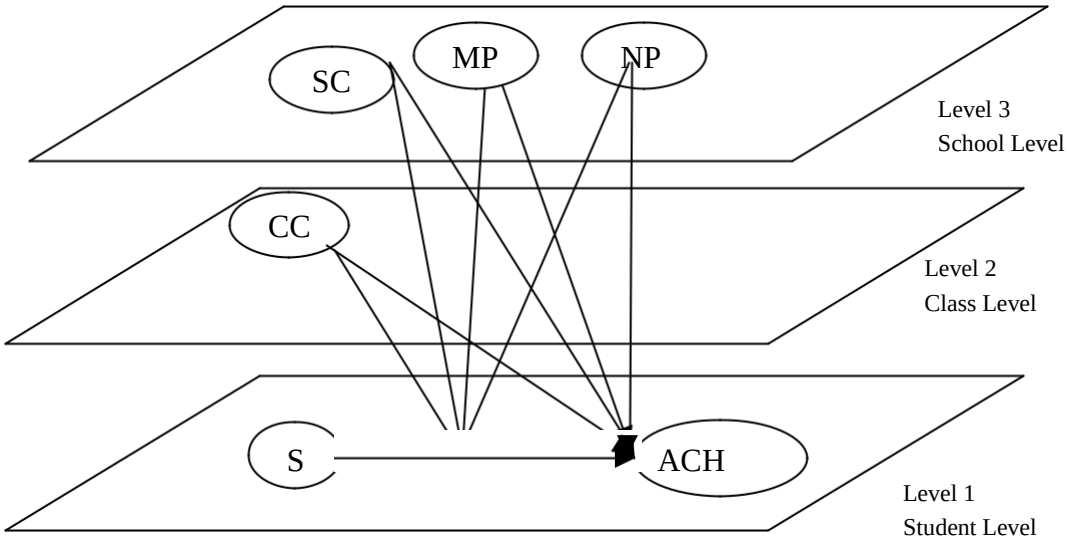


Figure 1. Hypothesised three-level hierarchical model for science achievement

ANALYSES

The multilevel models were built step-by step. The first step was to run a model without explanatory variables, which is also called the 'null model'. Thus null model was fitted to provide estimates of the variance components at each level (Raudenbush and Bryk , 2002). The null model can be stated in equation form as follows.

Level-1 model

$$Y_{ijk} = \pi_{0jk} + e_{ijk}$$

Level-2 model

$$\pi_{0jk} = \beta_{00j} + r_{0jk}$$

Level-3 model

$$\beta_{00k} = \gamma_{000} + u_{00k} \qquad (8)$$

where: Y_{ikj} is the science achievement of student i in class j in school k.

The second step undertaken was to estimate Type A effects in which student characteristics were added, thereby controlling for student intake. At this stage, a step-up approach was followed to examine which of the eight student-level variables (listed in Table 1) had a significant (at $p \leq 0.05$) influence on the outcome variable, ACH. Four variables (FOCC, EXPED, LIKSCI and PRIORACH) were found to be significant and therefore were included in the model at this stage. These four student-level variables were grand-mean-centred in the HLM analyses so that the

intercept term would represent the average ACH score for the students with average student characteristics. When a variable was centred around its grand mean, the zero value indicated its average value.

Table 1. Variables tested at each level of the hierarchy

Level	Variable code	Variable description
Level-1		(Student-level)
(S)	FOCC	Father's occupation (1=Unskilled labourer,. . . , 6= Professional)
	EXPOCC	Expected occupation (1=Unskilled labourer,. . . , 6= Professional)
	EXPED	Expected education (1=Year 10 and Below, . . . ; 6=Higher Degree)
	ACAMOT	Academic motivation (0=Lowest motivation, . . . , 40=Highest motivation)
	LIKSCH	Like school (0=Likes school least, . . . , 34=Likes school most)
	LIKSCI	Like science (1=Likes science least, . . . , 40=Likes science most)
	SELREG	Self regard (1=Lowest self regard, . . . , 34=Highest self regard)
	PRIORACH	Prior science achievement (0=Lowest score, . . . , 25=Highest score)
Level-2		(Class-level)
(CC)	CSIZE	Class size (8=Smallest, . . . , 39=Largest)
	FOCC_2	Average fathers' occupation at class-level
	EXPOCC_2	Average expected occupation at class-level
	EXPED_2	Average expected education at class-level
	ACAMOT_2	Average academic motivation at class-level
	LIKSCH_2	Average like school at class-level
	LIKSCI_2	Average like science at class-level
	SELREG_2	Average self regard at class-level
	PRIOR_2	Average prior science achievement
Level-3		(School-level)
(SC)	CSIZE_3	Average class size
	FOCC_3	Average fathers' occupation at school-level
	EXPOCC_3	Average expected occupation at school-level
	EXPED_3	Average expected education at school-level
	ACAMOT_3	Average academic motivation at school-level
	LIKSCH_3	Average like school at school-level
	LIKSCI_3	Average like science at school-level
	SELREG_3	Average self regard at school-level
	PRIOR_3	Average prior science achievement
(NP)	GOVT	Government school (0=Non-government; 1=Government)
	CATH	Catholic school (0=Non-Catholic; 1=Catholic)
	IND	Independent school (0=Non-Independent; 1=Independent)
	BOYS	Boys' school (0=Girls and Co-ed; 1=Boys only)
	GIRLS	Girls' school (0=Boys' and Co-ed; 1=Girls only)
	COED	Co-educational school (0=Boys only and Girls' only; 1=Co-ed)
(MP)	STREAM	Streaming in school (0=No streaming; 1=Streaming)
Outcome	ACH	Science Achievement (1 =lowest score….55=highest score)

The third step undertaken was to estimate Type B effects, which involved adding the classroom context and school context variables into the model using the step-up strategy mentioned above. At this stage, the Level-2 and Level-3 exploratory analysis sub-routines available in HLM 6 were employed for examining the potentially significant classroom and school context variables (as found in the output) in successive HLM runs. Following the step-up procedure, two classroom context variables (PRIOR_2 and CSIZE) were included in the model for the intercept. In addition, two cross-level interaction effects (between PRIORACH and FOCC_2 and between PRIORACH and LIKSCI_3) were included in the model.

The fourth step involved adding the significant non-malleable school policies and practices into the model using the Level-3 exploratory analysis sub-routine and the step-up strategy. At this stage, two cross-level interaction effects (between FOCC and GOV and between EXPED and

IND) were included in the model. In addition, the estimated coefficients for FOCC_2 were fixed at the school level because the reliability estimate of this coefficient was below 0.10.

The final step involved adding the malleable school policy into the model (STREAM). Estimates of fixed effects for Types A, B, X, and Z models have been given in Table 2.

The Type A model can be denoted as follows.

Level-1 model

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}FOCC_{ijk} + \pi_{2jk}EXPED_{ijk} + \pi_{3jk}LIKSCI_{ijk} + \pi_{4jk}PRIORACH_{ijk} + e_{ijk}$$

Level-2 model

$$\pi_{0jk} = \beta_{00k} + r_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$
$$\pi_{2jk} = \beta_{20k} + r_{2jk}$$
$$\pi_{3jk} = \beta_{30k} + r_{3jk}$$
$$\pi_{4jk} = \beta_{40k} + r_{4jk}$$

Level-3 model

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + u_{00k} \\
\beta_{10k} &= \gamma_{100} + u_{10k} \\
\beta_{20k} &= \gamma_{200} + u_{20k} \\
\beta_{30k} &= \gamma_{300} + u_{30k} \\
\beta_{40k} &= \gamma_{400} + u_{40k}
\end{aligned} \qquad (9)$$

The Type B model can be denoted as follows.

Level-1 model

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}FOCC_{ijk} + \pi_{2jk}EXPED_{ijk} + \pi_{3jk}LIKSCI_{ijk} + \pi_{4jk}PRIORACH_{ijk} + e_{ijk}$$

Level-2 model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}PRIOR\_2_{0jk} + \beta_{02k}CSIZE_{0jk} + r_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$
$$\pi_{2jk} = \beta_{20k} + r_{2jk}$$
$$\pi_{3jk} = \beta_{30k} + r_{3jk}$$
$$\pi_{4jk} = \beta_{40k} + \beta_{41k}FOCC\_2_{4jk} + r_{4jk}$$

Level-3 model

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + u_{00k} \\
\beta_{01k} &= \gamma_{010} + u_{01k} \\
\beta_{02k} &= \gamma_{020} + u_{02k} \\
\beta_{10k} &= \gamma_{100} + u_{10k} \\
\beta_{20k} &= \gamma_{200} + u_{20k} \\
\beta_{30k} &= \gamma_{300} + u_{30k}
\end{aligned}$$

$$\beta_{40k} = \gamma_{400} + \gamma_{401}LIKSCI\_3_{40k} + u_{40k}$$

$$\beta_{41k} = \gamma_{410} + u_{41k} \qquad (10)$$

The Type X model can be denoted as follows.

Level-1 model

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}FOCC_{ijk} + \pi_{2jk}EXPED_{ijk} + \pi_{3jk}LIKSCI_{ijk} + \pi_{4jk}PRIORACH_{ijk} + e_{ijk}$$

Level-2 model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}PRIOR\_2_{0jk} + \beta_{02k}CSIZE_{0jk} + u_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$
$$\pi_{2jk} = \beta_{20k} + r_{2jk}$$
$$\pi_{3jk} = \beta_{30k} + r_{3jk}$$
$$\pi_{4jk} = \beta_{40k} + \beta_{41k}FOCC\_2_{4jk} + r_{4jk}$$

Level-3 model

$$\beta_{00k} = \gamma_{000} + u_{00k}$$
$$\beta_{01k} = \gamma_{010} + u_{01k}$$
$$\beta_{02k} = \gamma_{020} + u_{02k}$$
$$\beta_{10k} = \gamma_{100} + \gamma_{101}GOV_{10k} + u_{10k}$$
$$\beta_{20k} = \gamma_{200} + \gamma_{201}IND_{20k} + u_{20k}$$
$$\beta_{30k} = \gamma_{300} + u_{30k}$$
$$\beta_{40k} = \gamma_{400} + \gamma_{401}LIKSCI\_3_{40k} + u_{40k}$$
$$\beta_{41k} = \gamma_{410} \qquad (11)$$

The Type Z model can be denoted as follows.

Level-1 model

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}FOCC_{ijk} + \pi_{2jk}EXPED_{ijk} + \pi_{3jk}LIKSCI_{ijk} + \pi_{4jk}PRIORACH_{ijk} + e_{ijk}$$

Level-2 model

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}PRIOR\_2_{0jk} + \beta_{02k}CSIZE_{0jk} + u_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$
$$\pi_{2jk} = \beta_{20k} + r_{2jk}$$
$$\pi_{3jk} = \beta_{30k} + r_{3jk}$$
$$\pi_{4jk} = \beta_{40k} + \beta_{41k}FOCC\_2_{4jk} + r_{4jk}$$

Level-3 model

$$\beta_{00k} = \gamma_{000} + \gamma_{001}STREAM_{00k} + u_{00k}$$
$$\beta_{01k} = \gamma_{010} + u_{01k}$$
$$\beta_{02k} = \gamma_{020} + u_{02k}$$
$$\beta_{10k} = \gamma_{100} + \gamma_{101}GOV_{10k} + u_{10k}$$
$$\beta_{20k} = \gamma_{200} + \gamma_{201}IND_{20k} + u_{20k}$$
$$\beta_{30k} = \gamma_{300} + u_{30k}$$
$$\beta_{40k} = \gamma_{400} + \gamma_{401}LIKSCI\_3_{40k} + u_{40k}$$
$$\beta_{41k} = \gamma_{410} \qquad (12)$$

VARIANCE EXPLAINED

The concept of variance explained is very common in multiple regression analysis. It gives the idea of how much of the variability of the dependent variable is accounted for by linear regression

on the predictor variables. The usual measure of the proportion of variance explained is the square multiple correlation, R^2. One way to approach this concept is to treat separately proportional reductions in the estimated variance components, σ^2, τ_0^2, and ϕ_0^2 at Level 1, 2, and 3 respectively as analogues of R^2 values at each level.

Variance components for the null model: σ_n^2, τ_{n0}^2, and ϕ_{n0}^2.

Variance components for the final model: σ_f^2, τ_{f0}^2, and ϕ_{f0}^2.

Proportion of variance explained at each level in the final model:

$$\text{At Level 1:} \quad R^2_1 = \frac{\sigma^2_n - \sigma^2_f}{\sigma^2_n}$$

$$\text{At Level 2:} \quad R^2_2 = \frac{\tau^2_{n0} - \tau^2_{f0}}{\tau^2_{n0}}$$

$$\text{At Level 3:} \quad R^2_3 = \frac{\phi^2_{n0} - \phi^2_{f0}}{\phi^2_{n0}} \qquad (13)$$

However, this approach can be somewhat problematic. It sometimes happens that adding explanatory variables increases rather than decreases some of the variance components. Therefore, it is possible to obtain negative values of R^2. Snijders and Bosker (1999) gave a suitable multilevel version of R^2 for the two-level model where the average class size was n_2 as follows:

$$\text{At Level 1:} \quad R^2_1 = 1 - \frac{\sigma^2_f + \tau^2_{f0}}{\sigma^2_n + \tau^2_{n0}}$$

$$\text{At Level 2:} \quad R^2_2 = 1 - \frac{\sigma^2_f / n_2 + \tau^2_{f0}}{\sigma^2_n / n_2 + \tau^2_{n0}} \qquad (14)$$

Equation (8) can be extended to a three-level model where on average each school consists of n_3 classrooms.

$$\text{At Level 1:} \quad R^2_1 = 1 - \frac{\sigma^2_f + \tau^2_{f0} + \phi^2_{f0}}{\sigma^2_n + \tau^2_{n0} + \phi^2_{n0}}$$

$$\text{At Level 2:} \quad R^2_2 = 1 - \frac{\sigma^2_f / n_2 + \tau^2_{f0} + \phi^2_{f0}}{\sigma^2_n / n_2 + \tau^2_{n0} + \phi^2_{n0}}$$

$$\text{At Level 3:} \quad R^2_3 = 1 - \frac{\sigma^2_f / (n_2 * n_3) + \tau^2_{f0} / n_3 + \phi^2_{f0}}{\sigma^2_n / (n_2 * n_3) + \tau^2_{n0} / n_3 + \phi^2_{n0}} \qquad (15)$$

Variance components presented in Table 3 were calculated using equation (15).

RESULTS

The Null Model: Differences Between Schools and Between Classes

The analysis was started by fitting the null model. This model provides estimates of the differences between students, between classes and between schools. The sum of these three

components is the total variance. It can be seen in Table 3 that for science achievement, 53.3 per cent (38.07/71.45) of the total variance is situated at the student level and another 46.6 per cent (33.34/71.45) of the total variance is located at the class level. These large components indicate that there are large differences between students and between classrooms. The percentage of the variance at the school level is very small (0.04/71.45=0.1%) which suggests that the schools are very similar to each other in terms of student achievement in science. In other words, the Level 3 intraclass correlation expressing the likeness of students in the same school is estimated to be 0.001, while the intraclass correlation expressing the likeness of students in the same classes and the same schools is estimated to be 0.47. Since most of the variance components at the school and class levels are situated at the class level, it is important to localise school performance to a specific classroom or grade level.

Type A Model: Adding Student Characteristics

At the student-level, the results in Table 2 show that Science achievement is directly influenced by Father's occupation (FOCC), Expected occupation (EXPED), Like science (LIKSCI) and Prior achievement (PRIORACH). When other factors were equal, students whose fathers had high status occupations outperformed students whose fathers had low status occupations. Students who aspired to pursue education to higher levels were estimated to achieve better when compared to students who had no such ambitions, while students who liked science were estimated to achieve better when compared to students who did not like science. In addition, students who had high prior achievement scores were estimated to achieve better than students who had low prior achievement scores.

Adding the student level variables to the model explains a large part of the differences between students (52.7 %), classes (69.9 %), and between schools (69.8 %) in science achievement. In other words, science achievement differences between schools and between classes were largely due to intake differences at the grade level under survey. The remaining differences between classes and between schools were indicators of the variance in Type A school effects and in Type A teaching effects. The residuals of schools and classes can be seen in Figure 2 and Figure 3 respectively, with little variability between schools.

Type B Model: Adding Classroom and School Contexts

From Table 2 it can be seen that at the class-level, Science achievement is directly influenced by Average prior achievement (PRIOR_2) and Class size (CSIZE). When other factors were kept equal, students in classes with high prior achievement scores were likely to achieve better when compared to students in classes with low prior achievement scores. Importantly, there was considerable advantage (in term of better achievement in science) associated with being in larger classes. Nevertheless, in interpreting the effects of class size, it needs to be recognised that 10 out of the 15 schools in these data had a streaming policy that involved placing high achieving students in larger classes and low achieving students in smaller classes for effective teaching. Therefore, the better performance of the students in larger classes in these data is not surprising. Students in the schools that implemented streaming policy achieved better in science.

Table 2. Final estimation of fixed effects

		Type A model			Type B model			Type X model			Type Z model		
Fixed Effects		Coefficient	S.E	p-value	Coefficient	S.E	p-value	Coefficient	S.E	p-value	Coefficient	S.E	p-value
Intercept	γ_{000}	28.52	0.25	0.000	28.34	0.31	0.000	28.31	0.32	0.000	27.17	0.50	0.000
STREAM	γ_{001}										1.62	0.59	0.017
PRIOR_2	γ_{010}				0.27	0.08	0.003	0.28	0.07	0.002	0.29	0.06	0.001
CSIZE	γ_{020}				0.28	0.06	0.001	0.29	0.07	0.001	0.30	0.06	0.000
FOCC,	γ_{100}	0.37	0.12	0.011	0.37	0.14	0.016	0.68	0.19	0.004	0.71	0.19	0.003
Interaction with GOV	γ_{101}							-0.50	0.22	0.044	-0.55	0.22	0.029
EXPED	γ_{200}	0.58	0.09	0.000	0.50	0.10	0.000	0.44	0.10	0.000	0.43	0.10	0.001
Interaction with IND	γ_{201}							0.54	0.28	0.072	0.66	0.29	0.041
LIKSCI	γ_{300}	0.14	0.01	0.000	0.15	0.01	0.000	0.14	0.01	0.000	0.15	0.01	0.013
PRIORACH	γ_{400}	0.97	0.05	0.000	0.93	0.04	0.000	0.94	0.04	0.000	0.94	0.04	0.000
Interaction with LIKSCI_3	γ_{401}				0.01	0.00	0.024	0.01	0.00	0.027	0.01	0.00	0.028
Interaction with FOCC_2	γ_{410}				0.07	0.02	0.020	0.05	0.02	0.023	0.06	0.02	0.026

Table 3. Variance Components

Model	Deviance	Number of Parameter Estimated	**Available**			**Explained (%)**			**Unexplained (%)**		
			Student (N=1984)	**Class** (K=71)	**School** (J=15)	**Student** (N=1984)	**Class** (K=71)	**School** (J=15)	**Student** (N=1984)	**Class** (K=71)	**School** (J=15)
Null model	13,078	4	38.07	33.34	0.04						
Prior Achievement	12,142	9	24.22	9.58	0.02	52.7	69.9	69.8	47.3	30.1	30.2
Type A Model	11,879	36	20.68	6.49	0.14	61.8	78.8	77.4	38.2	21.2	22.6
Type B Model	11,792	61	20.56	1.63	0.81	67.9	90.9	82.2	32.2	9.1	17.8
Type X Model	11,786	55	20.54	1.63	0.71	67.8	91.1	83.6	32.1	8.9	16.4
Type Z Model	11,783	56	20.54	1.74	0.31	68.4	92.0	88.7	31.6	8.0	11.3

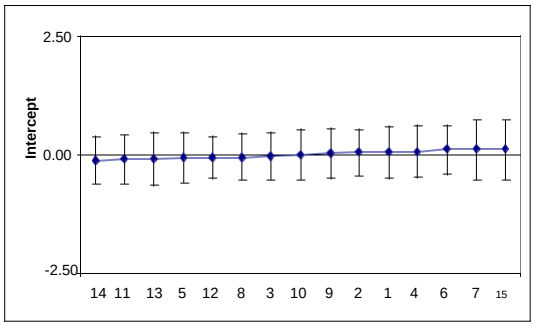


Figure 2. Type A school residuals

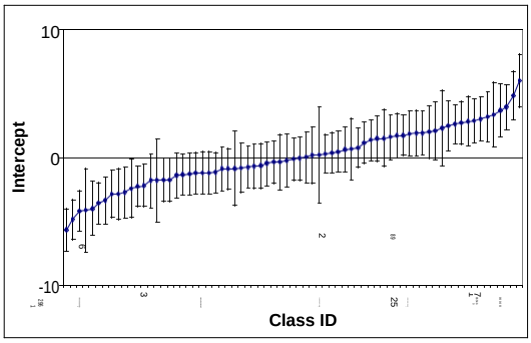


Figure 3. Type A class residuals

In Table 2 there are two significant cross-level interaction effects. These cross-level interaction effects are between (a) PRIORACH and FOOC_2 at Level 2 (class level); and (b) PRIORACH and LIKSCI_3 at Level 3 (school level). It can be seen in Figure 4 and Figure 5 that the effect of prior achievement is stronger in classes with higher status of fathers' occupation and in schools with higher level of liking science. Higher achieving students were better off in classes that had higher status of fathers' occupation as well as in schools with higher levels of liking science.

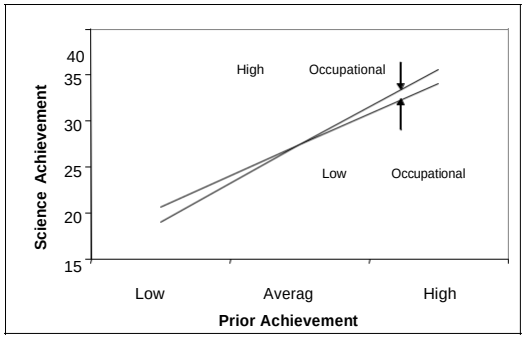


Figure 4. Impact of interaction effect of FOCC_2 and PRIORACH on Science Achievement at the classroom level

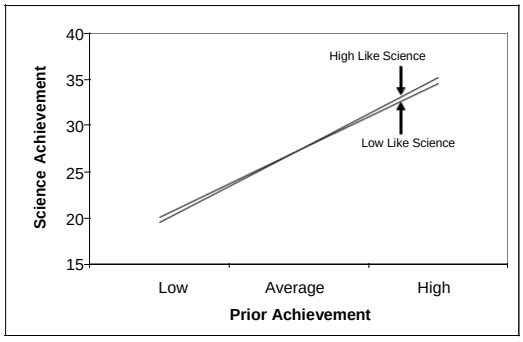


Figure 5. Impact of interaction effect of LIKSCI_3 and PRIORACH on Science Achievement at the school level

After controlling for student characteristics, class context and school context, the proportion of variance explained is increased by 9.1 per cent at the student level, 8.9 per cent at the class level, and 7.6 per cent at the school level. The residuals of 15 schools and 71 classes can be seen in Figure 6 and Figure 7, with an increase in variability between schools and a decrease in the variability between classes.

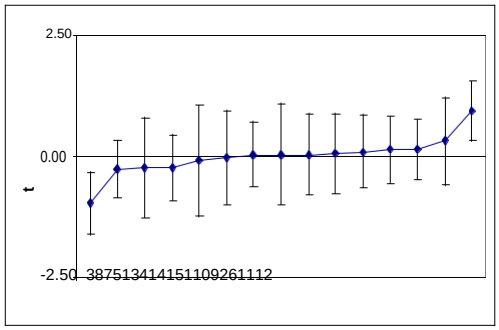


Figure 6. Type B school residuals

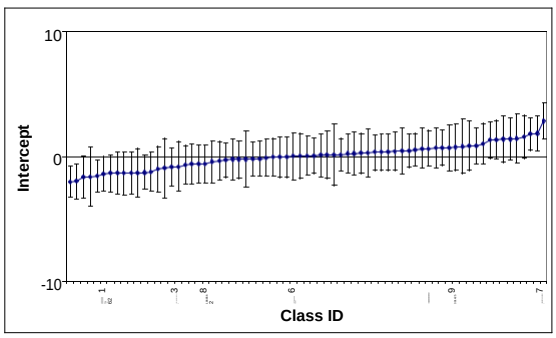


Figure 7. Type B class residuals

Type X Model: Adding Non-Malleable School Policies and Practices

When non-malleable policies were entered into the equations at Level 3, two additional interaction effects were found. These interaction effects included interactions between (a) FOCC and GOV and (b) EXPED and IND. From Figure 8 it can be seen that when other factors are equal, father's occupation had less impact in government schools than in non-government schools. In other words, students with high father's occupational status gained smaller advantage in government schools compared non-government schools.

Likewise, from Figure 9 it can be seen that students in independent schools achieve higher scores in science when they have high expected education. However, students with low levels of expected education have noticeably lower levels of achievement if they are in independent schools.

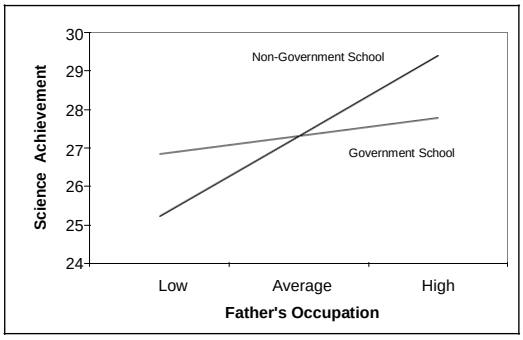


Figure 8. Impact of interaction effect of Government School and FOCC on Science Achievement at the school level

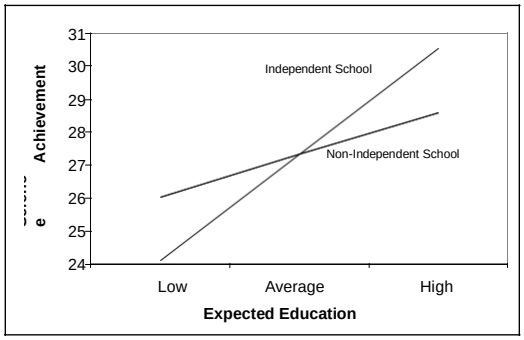


Figure 9. Impact of interaction effect of Independent School and EXPED on Science Achievement at the school level

After adding non-malleable policies and practices, only 16.4 per cent and 8.9 per cent of variance components at the school and class levels are left unexplained. The Type X residuals of 15 schools and 71 classes can be seen in Figure 10 and Figure 11 respectively.

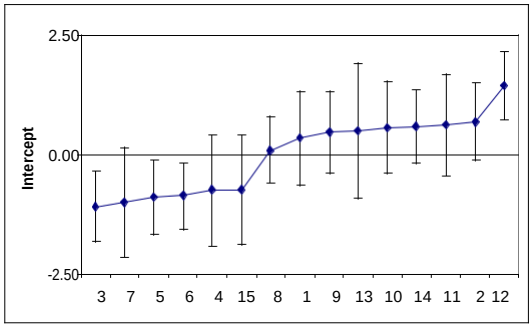


Figure 10. Type X school residuals

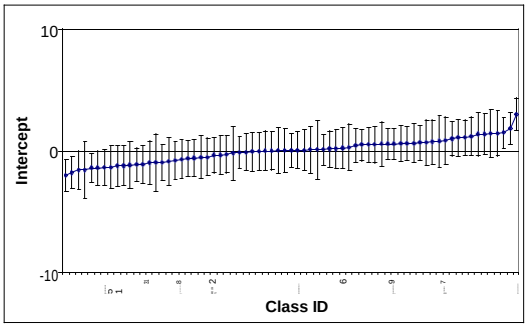


Figure 11. Type X class residuals

Type Z Model: Adding Malleable School Policies and Practices

At the School level, the results in Table 2 show that Science Achievement is also directly influenced by streaming policy (STREAM). Students in the schools that implemented streaming policy achieved better in science. In this model, only 31.6 per cent, 8.9 per cent and 11.3 per cent of variance components at student, class, and school levels are left unexplained. The Type Z residuals of 15 schools and 71 classes can be seen in Figure 12 and Figure 13 respectively.

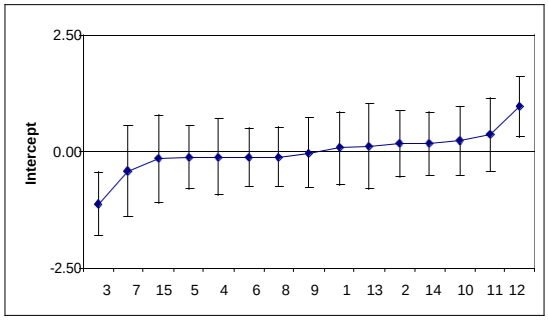


Figure 12. Type Z school residuals

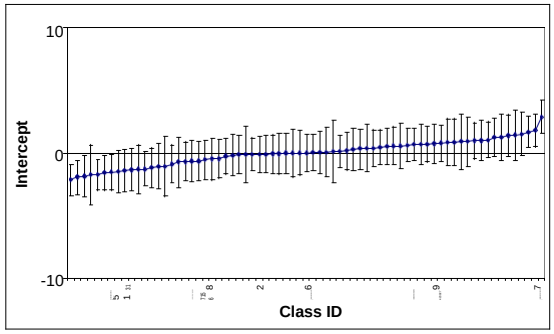


Figure 13. Type Z class residuals

Initially the differences between schools are very small as shown by the residuals of the Null model. After controlling for student characteristics, there are still no significant differences between schools. Adding classroom context and school context variables noticeably change the residuals for School 3 and School 12. Making allowance for additional non-malleable policies changed school residuals even further. However, after controlling for the significant malleable policy variable the average levels of performance for most schools are not significantly different from each other. School 3 and School 12 are the two schools that have noticeably lower and higher performance respectively. School 3 is significantly worse than other schools, but School 12 is significantly better than other schools after controlling for student characteristics, context variables as well as identified school policies and practices. These changes are noticeable from comparison of Figures 2, 6, 10, and 12 as well as from Figure 14, after allowance is made for the Type A, Type B, Type X and Type Z effects.

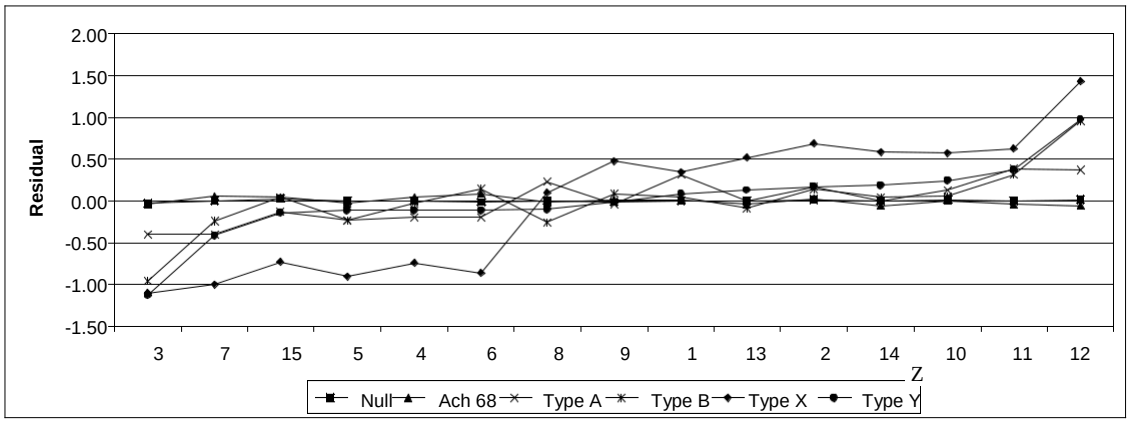


Figure 14. Types A, B, X and Z school residuals

CONCLUSIONS

This article is concerned with the statement that, student outcomes are only partially influenced by the school where they are enrolled. Other factors that have an impact on the student outcomes are student characteristics and context variables. In this study, Type A, Type B, Type X and Type Z effects are estimated by allowing for student background, class and school contexts, non-malleable school policies and malleable school policies respectively in successive regression equations.

The main effects reported from the analysis at the student level, indicate that in addition to prior achievement, it was the social psychological measures associated with the differences between students within classrooms that were having effects, namely, socioeconomic status, educational aspirations, and attitudes towards learning science. About 32 per cent of the variance between

students within classrooms is left unexplained, indicating that there are other student-level factors likely to be involved in influencing student achievement.

At the classroom level, about eight per cent of the total classroom variance or only 1.7 per cent of the total variance is left unexplained, with the average level of prior achievement of the class group having a significant effect. In addition, class size has a positive effect at this level on science achievement, with students in larger classes doing significantly better than students in smaller classes. This effect is likely to be confounded with factors associated with the qualities of the teachers assigned to teach the larger and the smaller class groups. Perhaps, this indicates the skill of the administration of the schools, particularly in those schools that adopt streaming practices to select the best teachers and allocate them to the higher performing students in larger classes. In addition, an interaction effect also reveals that the effect of prior achievement is stronger in classes with high status of fathers' occupation. High achieving students are better off in classes that have higher status of fathers' occupation.

At the school level of analysis, streaming directly explains some of the differences in levels of performance between schools in spite of the very small between school variance. The influences of the non-malleable variables involving school type and whether a school is single-sex or coeducational do not have direct effects on the educational outcome of science achievement, but they do have moderating or interaction effects. Thus whether the school is a Government or an Independent school interacts with Father's occupation and Expected education respectively to have small effects on the outcome variable. Nevertheless, it is this factor of school type that has had and continues to have a marked influence on changes in the provision of education in the Australian school systems. Unfortunately it is no longer possible to undertake research into this issue, because over-simplistic value added comparisons, that were made prior to the introduction of multilevel analytical procedures have contaminated this field of inquiry in Australia.

Two important findings emerge from this study. First, considerable variance is situated at the class level. Therefore in examining value added across schools, the class level can not be ignored. Otherwise, the class level variance components may be confounded with student level and school level variance components and lead to an overestimation of school differences. In educational effectiveness research, neglecting class context variables may lead to incorrect conclusions. Second, very little variance is left unexplained at the school and class levels to be accounted for by characteristics associated with school resources or by the direct effects of teachers. If the qualities of teachers are having effects they are associated with and are subordinate to the levels of initial achievement of the students whom they are assigned to teach, with high achieving students being placed in larger classes possibly with the better teachers.

However, the use of a value added approach in assessing school effectiveness is not without problems. There is still room for argument whether Type A, Type B, Type X or Type Z effects should be considered. Careful thought also needs to be given when considering which of the variables should be used in estimating Type A, Type B, Type X and Type Z effects. Moreover, how to obtain information on classroom effects is yet another question to address. Should longitudinal data rather than cross-sectional data be used? Apart from these problems which still need to be debated, the value added approach is providing a way to assess better the effectiveness and the accountability of schools as well as classrooms and teachers. Furthermore, it is clearly inappropriate to rank schools on terms of their performance and indeed to rank countries, without giving some consideration to these complex statistical problems. Nonetheless, research and scholarly debate needs to be carried out to develop a better understanding of the issues addressed.

REFERENCES

Ballou, D., Sanders, W. and Wright, P. (2004) Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-65.

Bryk, A.S. and Raudenbush, S.W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA.: Sage Publications.

De Fraine, B., Van Damme, J. and Onghena, P. (2002) Accountability of School and Teachers:what should be taken into account? *European Educational Journal*, 1(3), 403-428.

Hungi, N. (2003) *Measuring School Effects Across Grades.* Flinders University Institute of International Education Research Collection. No.6. Adelaide: Shannon Research Press.

Keeves, J.P., Hungi, N. and Afrassa, T. (2005). Measuring value added effects across schools: should schools be compared in performance? *Studies in Educational Evaluation*, 31(2-3), 247-266.

Meyer, R.H. (2000) Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief*, 3(3), 1-7.

Meyer, R.H. (2002) *Value-Added Indicators: Do They Make an Important Difference? Evidence from the Milwaukee Public Schools*, Paper presented at the Annual Meeting of the American Educational Research Association, April 2, 2002: New Orleans.

Raudenbush, S.W. and Willms J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307-335.

Raudenbush, S.W. and Bryk, A.S. (1994). Hierarchical linear models. In T. Husén and T.N. Postlethwaite (Eds.), *International Encyclopedia of Education: Research and Studies* (2nd ed., pp. 2590-2596). Oxford: Pergamon Press.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F. and Congdon, R.T. (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling* . Lincolnwood, IL: Scientific Software International.

Rowe, K.J. (2000) Assessment, league tables and school effectiveness: Consider the issues and 'let's get real'!, *Journal of Educational Enquiry*, 1(1), 73-98.

Rubin, D.B., Stuart, E.A. and Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage Publications.

Willms, J.D. and Raudenbush, S.W. (1989) A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, 2(6), 209-232.

International Education Journal, 2013, Volume 8, 189-206.
ISSN 1443-1475 © 2013

Percentage population plots: A proposition for a new strategy for data analysis in comparative education

Pawel Piotr Skuza
School of Education, Flinders University *pawel.skuza@flinders.edu.au*

One of the issues facing educational research workers today is the determination of the similarities and differences between countries and cultures in the factors that influence educational outcomes. The author of this article proposes a new approach to this problem. Usually when countries are compared, the complete student samples are taken into consideration. At the same time, there are differences between countries with regard to their educational policies towards high or low achieving students as well as the effects of different student characteristics on the educational outcomes for those groups. Population Percentage Plots propose a new way of comparing the effects across the whole range of performance of groups of students.

Cross-national research, secondary data analysis, science achievement, comparative education, high and low achieving students

INTRODUCTION

Since the mid-1960s there have been substantial developments in the provision of secondary and higher education not only in the developed countries of the world, but also in many developing countries. The marked expansion of secondary education and the growth of universities have placed heavy financial burdens both on the wealthy nations and also on those nations that have growing financial commitments for infrastructure development to cater for a rapidly expanding population. The demand for accountability from the education sector has led to the introduction of different international testing programs that have undertaken surveys to assess student achievement at different levels of education. The international testing programs conducted by the International Association for the Evaluation of Educational Achievement (IEA) and those conducted by the Organisation for Economic Cooperation and Development (OECD) have provided valuable information for comparisons of the average levels of achievement between countries.

Initially these testing programs were committed to undertaking multivariate analyses to identify the factors that influenced educational achievement both across countries and within countries. The between country comparisons were undertaken in a search for the factors that had strong effects on educational outcomes. These analyses were limited by a lack of appropriate statistical procedures that enabled the teasing out of the factors, which operated at different levels of analysis, namely the student, classroom, school, region and country. However, while these problems have gradually been resolved, an under emphasis has emerged on the accurate estimation of the mean level of achievement of a national education system, without concern for the spread of scores in educational achievement and attitudes and the modelling of factors that influence the variation in scores both within and between countries. Some countries have sought to undertake multilevel and multivariate analyses of the data collected at a particular level of education within a country, and to publish the results of such analyses separately in national

reports. However, there has been a noticeable absence of analyses that have examined change over time, across educational levels, and between kindred countries. As a consequence there has been little development of an understanding of the factors operating to influence student learning both within and across countries. Moreover, there has been little if any analyses conducted to examine how these factors have changed as a consequence of the marked expansion that has occurred in education. Of particular concern is that, there would seem to be a lack of interest in the performance of the very able students on whom the future of each nation must depend, particularly in the fields of science, mathematics and information and communications technology. In addition, there has been a lack of recognition of the significant role of such attitudes as perseverance, and interest in mathematics and science that required the accurate estimation of attitudinal data at the individual and sub-group levels.

It is the purpose of this article to develop a strategy for the examination of high and low performing sub-groups of students, both with respect to their achievement and their attitudes towards education, so that a greater understanding of the factors that influence both achievement and attitudes can be advanced.

Furthermore, this article is written at a time when each country is not only concerned with issues associated with "education for all" and "equality of educational opportunity", but is also very dependent on the development of talent, to support and advance the economic, scientific and technological development of the country.

CONCERN FOR THE DEVELOPMENT OF ABILITY

Theories of human learning indicate that students of the same age can be in different stages of cognitive development and can have different cognitive abilities. However, there seems to be a gap between the theories that highlight the variability among students and reports from achievement surveys that focus mostly on a whole national population and report national mean values and estimates of population, rather then sub-group, effects.

Furthermore, there is a second interesting issue. Different countries have different policies with regard to the allocation of more or less resources to help the higher achieving students. Two countries for example, that seem to differ in this matter, are Iran and the Republic of Korea. Iran took part in the Third International Mathematics and Science Study in 1999 and was classified in the 31st position out of 39 countries in Science achievement with an average scale score of 448 (3.8)[1] significantly below the international average of 488 (0.7) (Martin, 2000, p.32) . Similarly, in Physics, Iran was classified in the 33rd position out of 39 countries with an average scale score of 445 (5.7) significantly below the international average of 488 (0.9) (Martin, 2000, p.99). However, students from Iran, who took part in the International Physics Olympiads (IPhO) were at the top of the international competitors, and Iran's best student got first, eleventh, third, seventeenth, second and third position in IPhO in 1997, 1998, 1999, 2000, 2001, and 2002 respectively (IPhO websites). In contrast, the Republic of Korea was high in TIMSS 1999 Science and Physics achievement (fifth with a score of 549 (2.6), and fourth with a score of 544 (5.1)) and the best students from Korea also did well in the International Physics Olympiads and got 56th, third, ninth, third, 42nd and ninth position in successive IPhOs. It is also worthy of mention that the total number of participants each year in IPhO was between 265 and 350.

In summary, according to the TIMSS 1999 study, on average the general population of students from Iran did not perform well when compared with the top achieving countries. On the contrary,

[1] Standard errors appear in parentheses.

in the International Physics Olympiads participants from Iran were in the top level of rankings. However, in both cases the Korean students were high in the cross-national rankings. From this comparison of Korea and Iran, it might be concluded that both countries strongly supported their more able students, but for different reasons Iran's students on average were not as high as Korea's.

These two reported findings indicate that further research into the different levels of students' achievement may provide very interesting information. The main purpose of this article is to present some ideas, which may form the beginning of a new strategy of data analysis that allows comparison of the impact of particular factors on student achievement and attitudes across countries and across different performance subgroups.

There are several important questions that guided the development of the strategy discussed and that may influence the direction of this approach to analyses in the future. Consequently, this initial introduction to the proposed method is based on these questions. As a short introduction to the method it can be said that it applies the simple principle of ordered subgroups selected according to the level of performance to examine the change in the estimated metric regression coefficients for successive subgroups, and to detect a pattern of change in the metric regression coefficients as an indicator of the change in relationship across different performance subgroups.

Although some interesting patterns and conclusions are presented in this article, the proposed method clearly needs further development.

At this stage it should also be noted that all analyses were done using data from the first Science Survey within the Programme for International Student Assessment (PISA) conducted in 2000. However, Science was not the highly tested subject on the PISA 2000 surveys, that was on this occasion Reading, with a lesser emphasis on Mathematics and Science. Only on PISA 2006 was Science the highly tested subject, while Mathematics occupied this position in PISA 2003. In addition, it is of interest to note that programmable macros in SPSS were used extensively in the data analyses. The PISA dataset provides two kinds of estimates of science scores: a weighted likelihood estimates (WLE) and a set of plausible values that resulted from a conditioning process. Little is known about the effects of the WLE procedure on the spread of scores and the estimation of the achievement levels of high performing students. In all analyses presented in this article only WLE estimates are used, therefore it has to be pointed out that any findings from this article relate to these estimates of achievement outcomes for samples collected in the PISA survey for each country under investigation.

Because of the novelty of the proposed method and the necessity for further development, collection of SPSS and Excel files, which were used, can be readily available for verification and request by e-mail (pawel.skuza@flinders.edu.au).

ELABORATION OF THE PROBLEM AND AN INTRODUCTION TO THE METHOD.

Question 1: Do the same relationships hold across different student performance levels as apply across the total student performance group for each country's sample?

A graph for different subgroups of students from Australia's PISA 2000 sample on the horizontal axis is presented in Figure 1. The line goes from a group of the top five per cent achievers to the top ten per cent and so on through the 100 per cent and bottom 90 per cent and to the bottom five per cent.

The vertical axis shows an unstandardised or metric regression coefficient (b), which was calculated between the so-called 'Warm estimate score' or 'weighted likelihood estimate' (WLE) for Science achievement and Sex of student for each achieving subgroup. The variable Sex of student was coded '1' for female, '2' for male. The standard deviations of scores for successive

subgroups change markedly across sub-groups, and as a consequence, correlations and standardized regression coefficients cannot be compared across subgroups, although metric regression coefficients can be meaningfully compared.

Obviously, it is possible that an unstandardised regression coefficient between Science achievement regressed on Sex of student can be close to zero and sex does not significantly influence science achievement when considered for the whole student sample. However, there is a significant positive relationship when the sample is gradually restricted to the higher achieving students and a significant negative relationship for lower achieving students. This fact, about boys doing better than girls in the higher achieving groups and worse when considering the lower achievers, is not unexpected.

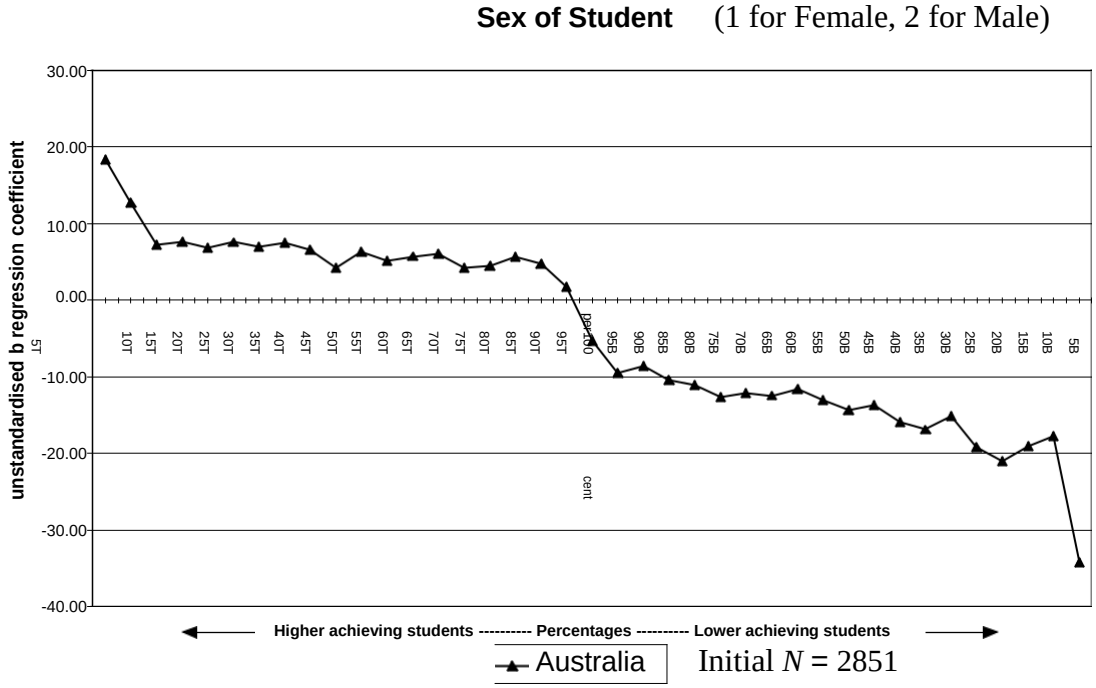


Figure 1 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Sex of student for different Science achievement subgroups from Australia (Source file A1.1).

Figure 2 and Table 1 are presented below to assist with an explanation of the analysis carried out and the graph drawn in Figure 1. In Table 1 the results from Excel are presented with all b regression coefficients, standard errors and significance tests for Australia for the regression of Science achievement on the variable Sex of student for all subgroups examined. Although, more graphs like that in Figure 1 are presented in this paper, all additional information like that in Table 1 is not included but is available online in the appropriate Excel files through an AutoFilter option. In order to draw Figure 1 it was necessary to calculate 39 metric regression coefficients for successive achievement subgroups.

The syntax file with macros in it, enabled the performance of this task to be carried out efficiently. Without providing great detail, it would be of value to describe briefly the general construction of the syntax file that was used to generate the regressions coefficients. This syntax file, throughout a series of loops, allowed for the selection from the PISA data file cases for the required countries and for the required percentage groups and finally for the required variables. Calculated metric regression coefficients together with their standard errors and significance tests were merged and sent to Excel files. At each stage of developing the syntax file, the cross tests

were undertaken to ensure that the obtained coefficients were correct. So, for example, Figure 2 shows the output from SPSS when a regression coefficient was calculated as a cross -test without using macros. The unstandardised regression coefficient ($b = 4.29$) from Figure 2 is equal to the value of the relevant point (T50) in Figure 1.

Table 1 Part of the file A1.1 with data, which were used to generate Figure 1 and with standard errors and significance tests (T-values are also reported) associated with the regression of Science achievement on Sex of Students

	b	SE	T	Sig.		b	SE	T	Sig.
Top 5	18.33	7.19	2.55	0.01	**100 percent**	-5.3	3.69	-1.44	0.15
Top 10	12.81	5.38	2.38	0.02	Bottom 95	-9.55	3.43	-2.78	0.01
Top 15	7.27	4.51	1.61	0.11	Bottom 90	-8.63	3.36	-2.57	0.01
Top 20	7.61	3.94	1.93	0.05	Bottom 85	-10.39	3.33	-3.12	0
Top 25	6.86	3.56	1.93	0.05	Bottom 80	-11.04	3.33	-3.32	0
Top 30	7.57	3.3	2.29	0.02	Bottom 75	-12.61	3.32	-3.80	0
Top 35	6.95	3.12	2.23	0.03	Bottom 70	-12.1	3.34	-3.62	0
Top 40	7.54	3.01	2.50	0.01	Bottom 65	-12.46	3.35	-3.72	0
Top 45	6.6	2.94	2.24	0.02	Bottom 60	-11.57	3.39	-3.41	0
Top 50	4.29	2.89	1.48	0.14	Bottom 55	-13.03	3.44	-3.79	0
Top 55	6.38	2.86	2.23	0.03	Bottom 50	-14.3	3.51	-4.07	0
Top 60	5.17	2.85	1.81	0.07	Bottom 45	-13.67	3.59	-3.81	0
Top 65	5.72	2.86	2.00	0.05	Bottom 40	-15.96	3.69	-4.33	0
Top 70	6.04	2.88	2.10	0.04	Bottom 35	-16.9	3.85	-4.39	0
Top 75	4.31	2.92	1.48	0.14	Bottom 30	-15.19	4.06	-3.74	0
Top 80	4.56	2.98	1.53	0.13	Bottom 25	-19.19	4.32	-4.44	0
Top 85	5.64	3.06	1.84	0.07	Bottom 20	-21	4.79	-4.38	0
Top 90	4.76	3.15	1.51	0.13	Bottom 15	-19.04	5.6	-3.40	0
Top 95	1.75	3.29	0.53	0.59	Bottom 10	-17.79	6.96	-2.56	0.01
100 percent	-5.3	3.69	-1.44	0.15	Bottom 5	-34.25	9.26	-3.70	0

Initial $N = 2851$ Boys scored 2, Girls scored 1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	593.21	4.60		128.92	.000
	Sex - Q3	4.29	2.89	.039	1.48	.139

a. Dependent Variable: Warm estimate in Science (WLE)

Figure 2 Part of the output from SPSS generated without using the macro for the top 50 per cent of Australia's sample and variable Sex of Student

Another graph is shown in Figure 3 where an unstandardised regression coefficient is plotted for different achievement subgroups. Again there is an interesting relationship for the high-achieving

students. In this case the variable, that is analysed, is 'Sense of belonging' and in order to provide a better understanding of it, the quotation from the *PISA 2000 Technical Manual* is presented[2]. In this example, as shown in Table 2, negative and significant regression coefficients are shown so, for example, in the top ten per cent of students those who did not have a strong sense of belonging to the school performed better than those students with a higher sense of belonging.

Table 2 Tests of significance for variable Sense of belonging and higher achieving subgroups of students in Figure 3

Top 5%	Top 10%	Top 15%	Top 20%	Top 25%	Top 30%	Top 35%	Top 40%	Top 45%	Top 50%	Top 55%
0.04	0.10	0.05	0.10	0.02	0.06	0.12	0.06	0.08	0.24	0.16

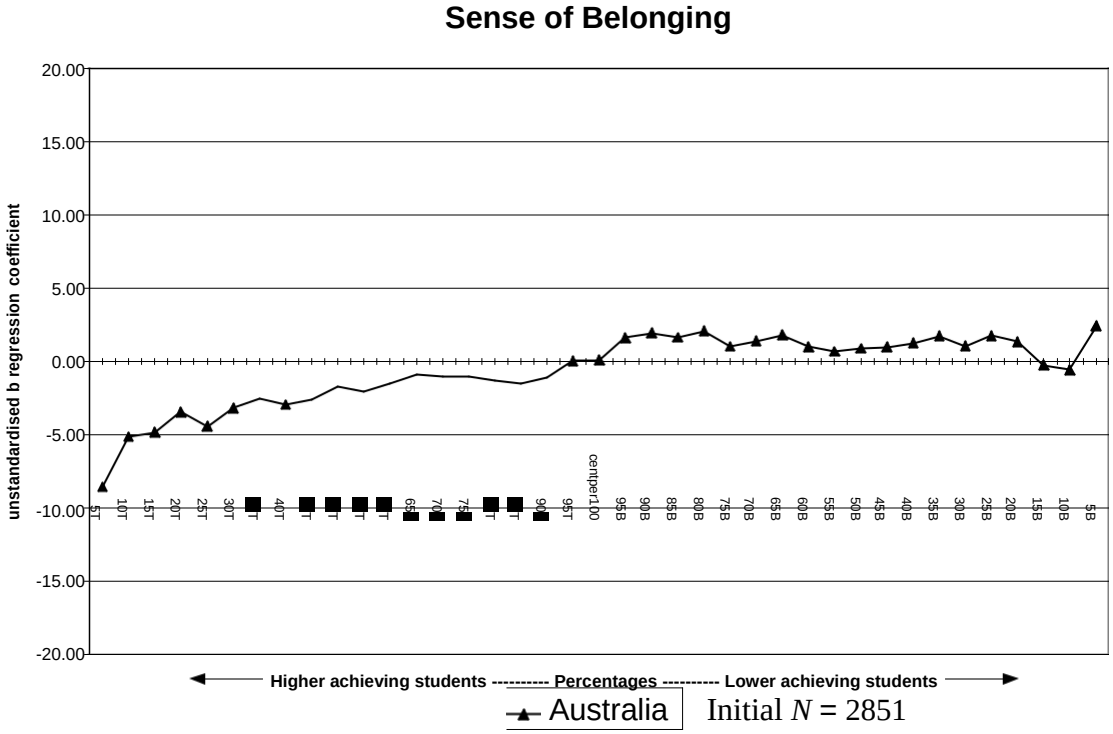


Figure 3 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable 'Sense of belonging' for different Science achievement subgroups from Australia (Source file A1.2). (NB Vertical scale is set from –20 to 20 because of later comparison between different countries)

The advantage of the graphs like those in Figure 1 or Figure 3, which can be called 'Percentage Population Plots' (PP -plots), is that successive unstandardised regression coefficients are presented to reveal a pattern associated with the successive levels of achievement of the subgroups of science students ranging from a small group (top 5 %) on the left-hand side of the

2 The PISA index of sense of belonging was derived from students' reports on whether their school is a place where they: feel like an outsider, make friends easily, feel like they belong, feel awkward and out of place, other students seem to like them, or feel lonely. A fourpoint scale was used with response categories: strongly disagree, disagree, agree and strongly agree. Scale scores are standardised Warm estimates where positive values indicate more positive attitudes towards school and negative values indicate less positive attitudes towards school.

(Adams and Wu, 2002, p. 229)

graph to the small group (bottom 5 %) on the right -hand side of the graph. While the extreme groups in the graphs are relatively small (about 140 cases in the case of the Australian sample) and may have sizable errors associated with the metric regression coefficients, the regression coefficients in the middle section of the graph have much smaller errors, since they are associated with large student groups. The statistical significance of the unstandardised regression coefficients, while estimated in Table 1 under the assumption of a simple random sample, does not take into consideration the changing design effects for each estimate, that necessarily increase the standard errors and increase p values associated with the tests of significance. However, the pattern of the graphs drawn in Figures 1 or 3 is highly informative, although the estimates of the unstandardised regression coefficients in the tails of graphs sometimes clearly indicate instability in the estimation procedure involved in the PP-plots.

The two above examples of graphs are presented merely to introduce the idea that there are differences between successive achievement groups in their regression relationships for particular variables. Because the main aim of this article is to introduce the PP -plots and show some possibilities of using these graphs, the issue of explaining why patterns of particular shapes occur is not developed any further at this stage.

Question 2: Does the same relationship hold across different student performance levels when comparing countries?

It is a well-known fact that there are differences between countries in the extent to which some factors influence student achievement; unfortunately most of the published findings report relationships for a whole population. It is mentioned above that the information available for Iran and Korea seems to show that both countries place great importance on supporting their more able students to take part in the International Physics Olympiads. On the contrary, there is a marked difference in the mean level of performance of the students between these two countries. Similarly there can be situations in which, for other pairs of countries, the students on average perform at the same level, but there are marked differences in the performance of the lower achieving students. Therefore, it is interesting to compare how different variables relate to Science achievement across different countries and across different achievement subgroups.

Family wealth

In Figure 4 the PP-plot is presented for six countries in which a PISA variable Family wealth[3] is argued to have a positive influence on student Science achievement, although with different values for the starting values of population estimates for the different countries. For five of the countries the path is symmetrically declining when moving towards higher and lower achieving students groups, except for Japan for which for almost all achieving subgroups Family wealth is not related to the Science achievement scores. For all countries shown in Figure 4, the PP -plots are roughly symmetrical when the left half of the graph is compared with the right half. Interestingly, this symmetrical relationship is not always shown for all countries. In Figure 5 a group of countries are presented, for which regression coefficients calculated between Family wealth and Science achievement scores are higher when moving towards better achieving students than when moving towards lower achieving subgroups of students. Moreover, in Figure 6 the opposite situation is shown.

---

3 The PISA index of family wealth was derived from students' reports on: (i) the availability in their home of a dishwasher, a room of their own, educational software, and a link to the Internet; and (ii) the numbers of cellular phones, televisions, computers, motor cars and bathrooms at home. Scale scores are standardised Warm estimates, where positive values indicate more wealth-related possessions and negative values indicate fewer wealth-related possessions.

(Adams and Wu, 2002, p. 224)

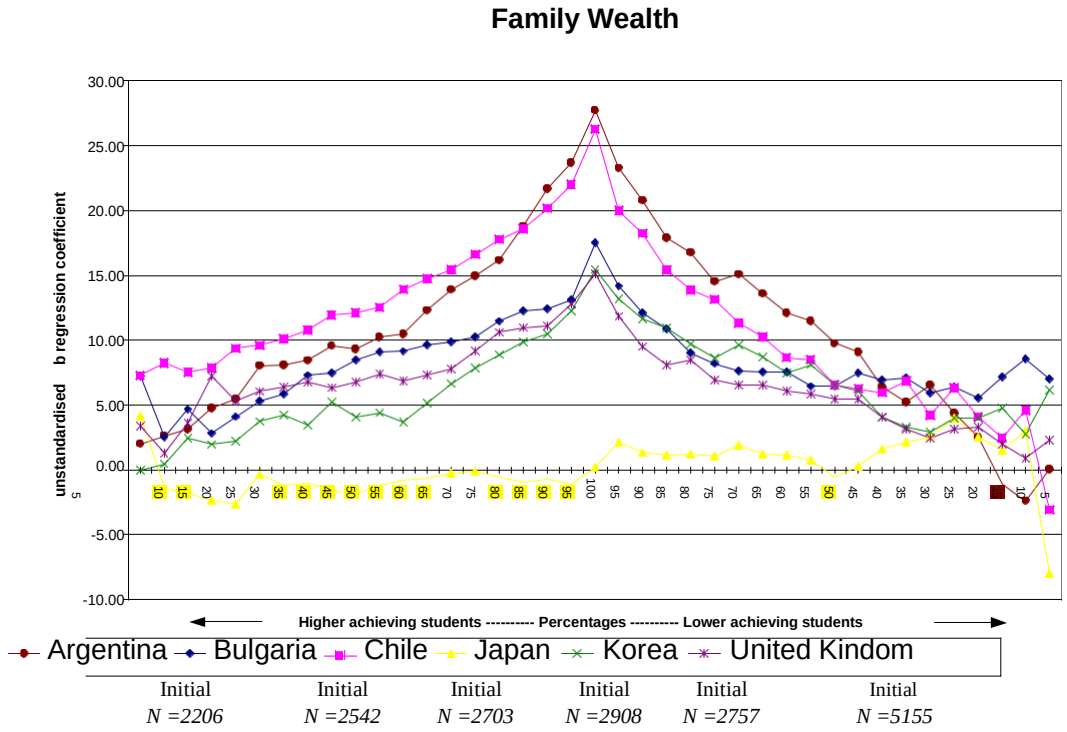


Figure 4 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Family wealth for different Science achievement subgroups from six countries: Argentina, Bulgaria, Chile, Japan, Korea, and United Kingdom (Source file A1.3).

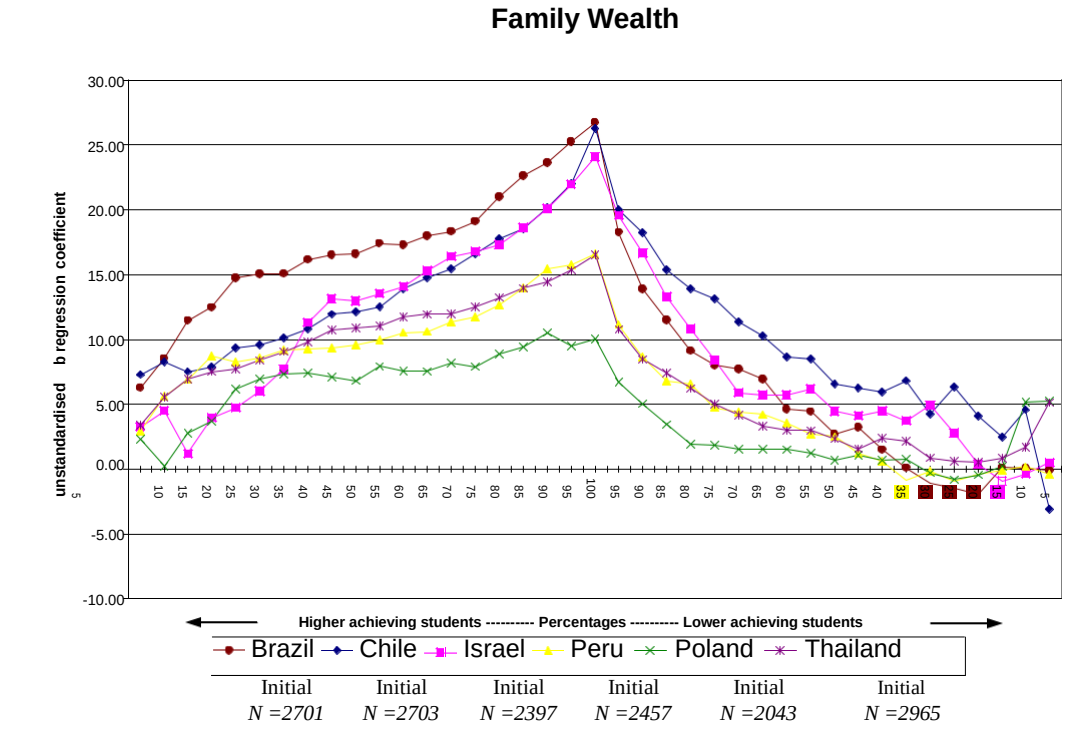


Figure 5 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Family wealth for Science achievement subgroups from six developing countries: Brazil, Chile, Israel, Peru, Poland, and Thailand (Source file A1.3).

Clearly, when comparing Figure 5 and Figure 6, countries seem to be grouped with respect to their level of development. In the case of the more developed countries in Figure 6 there is a change in the regression coefficient from positive to zero or even to negative with movement towards the high achieving subgroups of students. On the one hand it can be argued that such a pattern occurs because in developed countries there is free education with small differentiation in teaching quality between schools, so that the students who want to study Science, have plenty of opportunities to do so, regardless of their family wealth. Moreover, the negative sign of the regression coefficients, which indicate that students from richer families obtain lower scores compared to students from poorer families, may be because richer students' parents do not encourage their children to study science, preparing them to study law, economics and commerce. Alternatively, it may be likely that a career in science related fields provides greater possibilities for upward social mobility that is sought by students from poorer families.

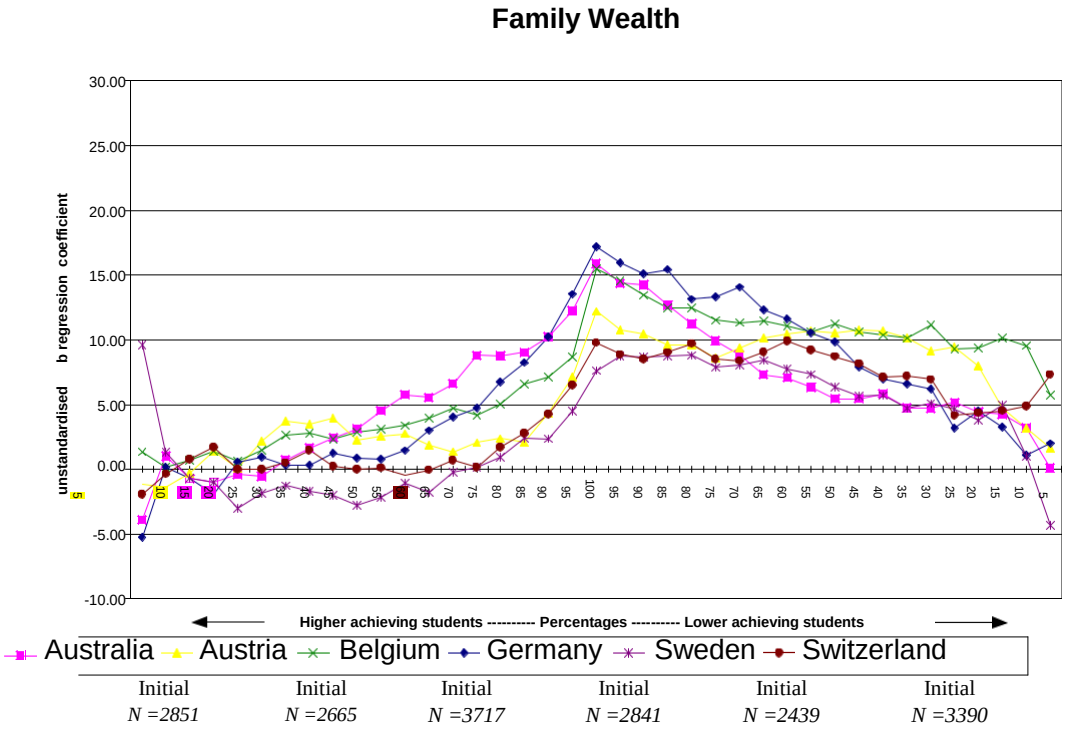


Figure 6 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Family wealth for different Science achievement subgroups from six developed countries: Australia, Austria, Belgium, Germany, Sweden, and Switzerland (Source file A1.3).

On the other hand in Figure 5 an opposite relationship is shown for developing countries in which Family wealth is positively related to Science achievement for higher achieving students. This seems to indicate that there are greater career rewards in scientifically based occupations (e.g. Medicine) that are very attractive to students from wealthy homes in developing countries.

Scatter plots and the meaning behind PP-plots

It is useful to add an additional explanation that may help to provide a better understanding of the meaning behind PP-plots. In a sense a PP-plot provides more detailed information about the shape of the scatter plot that shows the relationships between Science achievement and, in the case considered above, Family wealth. In Figures 5a and 6a the scatter plots for two countries Brazil and Germany are presented in relation to their PP-plots in Figures 5 and 6 respectively for the complete sample. It can be seen in the case of Brazil that the regression line does not change

substantially when compared to the complete sample, when the low achieving students are dropped. This shows that family wealth relates positively to the students' achievement even for the better students. However, the regression line becomes flatter if the high achieving students are deleted. That is clearly seen in the PP-plot in Figure 5 as well. When looking at the scatter plot for Germany (Figure 6a), it is seen to correspond to the PP-plot pattern. For example, for the high achieving students it is seen, that the regression line is flatter.

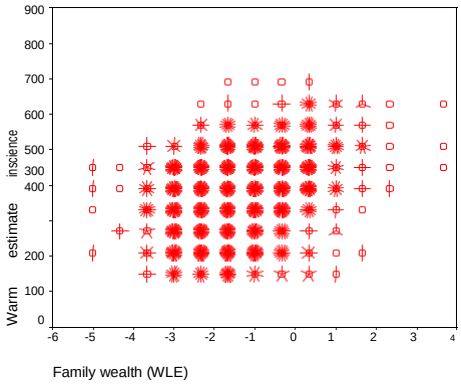


Figure 5a Scatter plot with relationship between Family wealth and Science achievement for Brazil

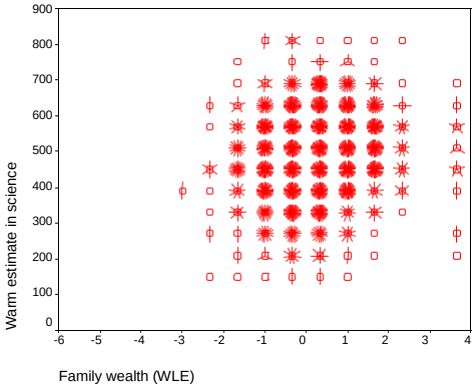


Figure 6a Scatter plot with relationship between Family wealth and Science achievement for Germany

Quite often in a research situation it is difficult to explain the relationships lying behind one particular regression coefficient in an estimated path model, even when many well established statistical methods are available. In this article a more general approach is presented for comparing regression coefficients for many countries and achievement subgroups. Perhaps it makes the task of examining the estimated relationships even more difficult. Many questions have to be asked even before starting, for example: How well does the Family wealth PISA variable reflect wealth in very poor countries? Therefore, explanations advanced are often highly speculative. Nevertheless, the main purpose of this article is to introduce the PP- plots and to examine such relationships further in order to understand better the meaning of the graphs. However, Figures 4, 5, 6 seem to address interesting so- called 'big picture' issues. The exception in the case of Japan shown in Figure 4 is of considerable interest.

Co-operative learning

Another example of an interesting between countries grouping is shown when analysing the PP-plots for the variable, Co-operative learning[4]. PP-plots for two groups of countries are presented in Figures 7 and 8. For the countries in Figure 8 values of the regression coefficients are restricted to the range –5 and 5, showing that a self-perceived view about Co-operative learning does not relate to Science achievement. This may be due to the very limited use of co-operative learning techniques in educational curricula within those countries shown on Figure 8. On the contrary, Figure 7 it can be seen that in subgroups with lower achieving students, those students, who have a preference for co-operative learning, are doing better in science.

4 The PISA index of co-operative learning was derived from student reports on the four items in Figure 64. A four-point scale with the response categories disagree, disagree somewhat, agree somewhat and agree was used. For information on the conceptual underpinning of the index, see Owens and Barnes (1992). Scale scores are standardised Warm estimates where positive values indicate higher levels of self-perception of preference for co-operative learning and negative values lower levels of self-perception of this preference. How much do you disagree or agree with each of the following? I like to work with other students, I learn most when I work with other students, I like to help other people do well in a group, It is helpful to put together everyone's ideas when working on a project. (Adams and Wu, 2002, p. 237)

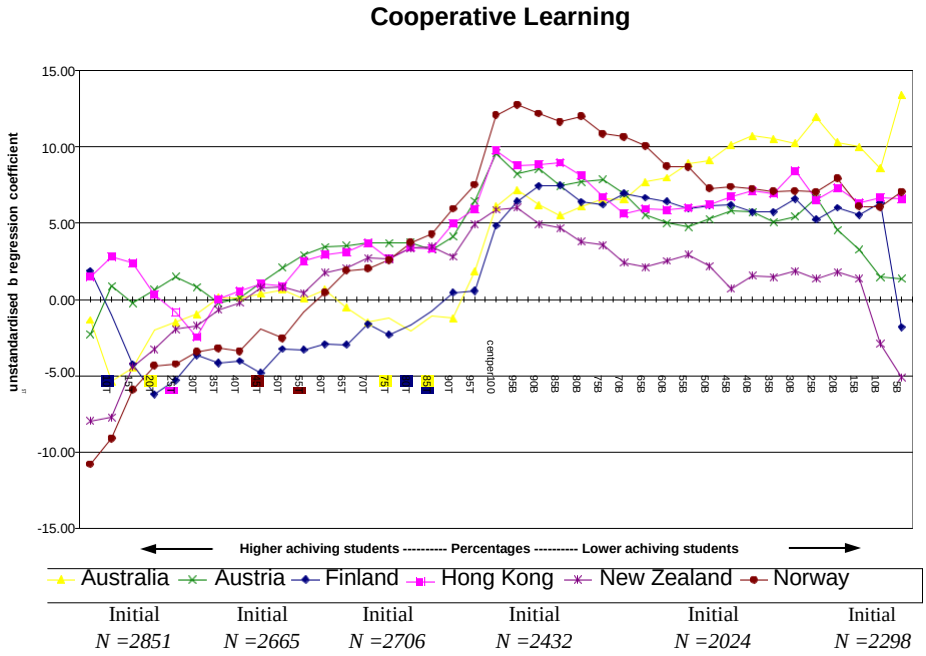


Figure 7 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Co-operative learning for different Science achievement subgroups from six developed countries: Australia, Austria, Finland, Hong Kong (China), New Zealand, and Norway (Source file A1.4).

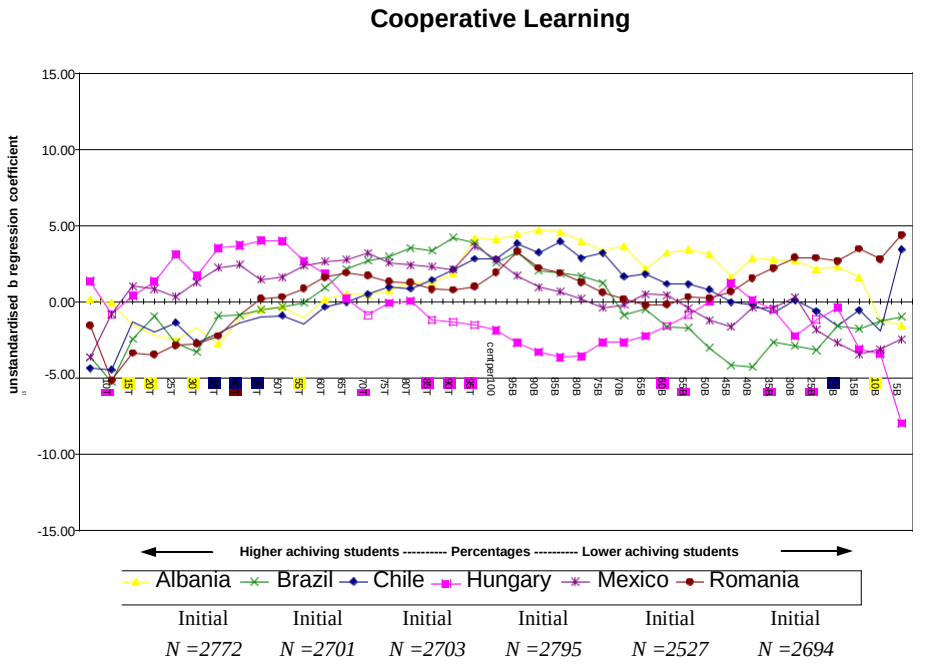


Figure 8 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Co-operative learning for different Science achievement subgroups from six developing countries: Albania, Brazil, Chile, Hungary, Mexico, and Romania (Source file A1.4).

Question 3: Can PP-plots help in detecting which variable is more generally intra-student or individually based and which is more generally inter-student or culturally based for all student achievement subgroups or for particular student achievement subgroups?

On the one hand, when looking at the PP-plots in Figures 4, 5 and 6, there are differences with regard to the extent to which the variable Family wealth is positively related to the Science achievement scores at the whole population level. There are even greater differences in the values and signs of the regression coefficients when considering the different achievement subgroups. Consequently it is meaningful to conclude that the relationship between Family wealth and Science achievement may depend on the kind of culture the students come from. On the other hand, in Figure 9 the PP- plots for the variable Sex of Students from all countries in PISA 2000 are presented. It may be argued that, although there are differences between the PP-plots, a general pattern seems to hold except in the left and right tails of the PP-plots among the most able and least able students. Therefore, it may be said that the way this variable relates to Science achievement outcomes is less culturally based and more individually based, or alternatively there is very little variability in the gender based societal differences between the countries involved and as a consequence in expected achievement in Science.

Probably because all the international surveys of students' knowledge and Science achievement, that have been conducted so far, have collected data about students' sex, similar PP-plots could be generated and may give clues to support or disprove the above assumption.

It is likely to be very useful from the policy makers' point of view, to know which factors influence students' Science achievement and are not due to cultural impact.

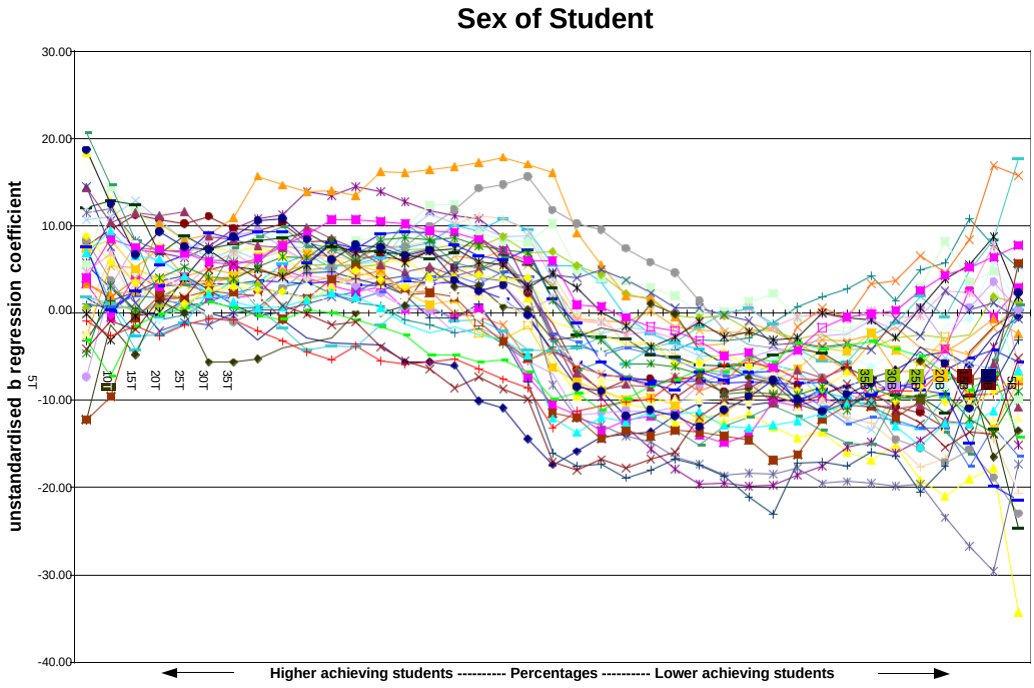


Figure 9 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Sex of students for different Science achievement subgroups and for 42 countries (Source file A1.1).

Another general conclusion can be drawn from Figure 10. For 30 countries out of 33 it can be shown that the values of the regression coefficients of Science achievement regressed on

Academic self-concept[5] for the higher achieving subgroups of students are greater than those on the right-hand side of the PP-plots. This is not unexpected, because many recorded findings support the proposition that students with higher academic self-concept are achieving at a higher level than those with lower academic self-concept. Interestingly, the three countries which break the pattern are Brazil, Romania and Thailand, and the three with the highest PP-plots are Australia, Denmark and Sweden.

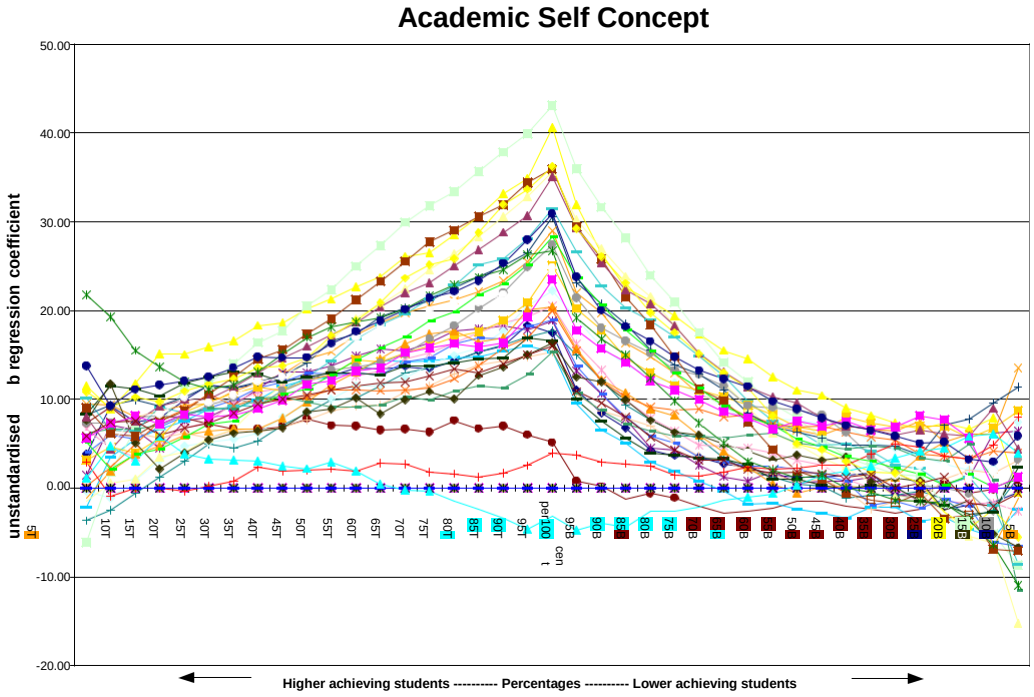


Figure 10 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable Academic self-concept for different Science achievement subgroups from 33 countries (Source file A1.5).

OTHER POSSIBILITIES ARISING FROM THE USE OF PP-PLOTS.

In a similar way, already developed syntax and macros with some small adjustments can be readily used with datasets from previous international surveys like IAEP, TIMSS, TIMSS-Repeat or PISA 2003. Obviously not all of the datasets collected are similar to the PISA 2000 set with respect to additional information from students, but for some variables, (for example, Sex of student) it is possible to generate PP-plots and to compare them with each other.

Moreover, generated in the way proposed, datasets with unstandardised regression coefficients can also be used in the meta-analyses. This possibility has not been developed further at this stage, but the regression coefficients from PISA2000 for whole national samples were used as a

---

[5] The PISA index of academic self-concept was derived from student responses to the items in Figure 68, which gives item parameters used for the weighted likelihood estimation. A four-point scale with the response categories disagree, disagree somewhat, agree somewhat and agree was used. For information on the conceptual underpinning of the index, see Marsh, Shavelson and Byrne (1992). Scale scores are standardised Warm estimates where positive values indicate higher levels of academic self-concept and negative values, lower levels of academic self-concept.
How much do you disagree or agree with each of the
following? I learn things quickly in most school subjects.
I'm good at most school subjects.
I do well in tests in most school subjects.

(Adams and Wu, 2002, p. 238)

base dataset for a Hierarchical Cluster Analysis. Interesting and perhaps to be expected, clustering of the countries on the basis of 20 variables (listed in Table 3) were available for all 42 countries and is presented in Figure 11. Three missing coefficients in the total dataset of the 837 coefficients were replaced with mean values. From Figure 11 it can be concluded that those 20 factors influence the Science achievement in a similar way for countries that are clustered closely together. A particularly strong and separate cluster is formed by Bulgaria, Czech Republic, Hungary and Poland. This cluster is an example, that may not have been predicted in advance, but when observed is highly meaningful and of considerable interest.

Table 3 The list of variables used in Hierarchical Cluster Analysis together with descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Sex	42	-17.40	16.03	-2.28	7.62
Mother international social and economical index	42	.50	1.95	1.30	.35
Father international social and economical index	42	.41	2.04	1.31	.41
Student self-expected international social and economical index	42	.24	2.36	1.51	.54
In. Socio-Econ. Index of father or mother	42	.34	2.07	1.31	.40
Father ISCED qualification	42	6.61	36.67	15.88	6.61
Mother ISCED qualification	42	5.63	35.01	16.93	6.39
Parental Academic interest (WLE)	42	4.25	26.53	15.64	5.58
Patental Social interest (WLE)	42	.81	19.07	8.75	4.43
Family educational support (WLE)	42	-17.49	8.78	-8.44	5.07
Family wealth (WLE)	42	-6.33	27.73	13.78	7.79
Home educational resources (WLE)	42	7.13	33.61	19.07	5.96
Cultural activities of students (WLE)	42	-3.84	30.39	14.75	7.95
Cultural possession of the family (WLE)	42	-.50	29.75	19.64	6.36
Time spent on homework (WLE)	42	-4.80	26.74	12.60	8.41
Teacher support (WLE)	42	-12.01	13.01	.52	5.66
School disciplinary climate (WLE)	42	-17.19	11.33	-6.05	5.83
Teacher-student relationship (WLE)	42	-8.44	14.91	3.43	7.77
Achievement press (WLE)	42	-14.05	11.33	-2.59	6.76
Sense of belonging (WLE)	42	-1.86	19.61	7.39	6.10

There is another possibility, although also not as yet developed, that may involve using the PP-plots to provide a way to group and classify countries. For example, the areas under the PP-plot curves for separate halves can be calculated and divided by each other. In this way an index for each country can be generated. In the case of the variable Family wealth such an index may provide information about how egalitarian particular countries are with respect to Science education. In the case of Academic self-concept such an index may help in the investigation of the degree to which academic self-confidence promotes higher achieving students compared to lower achieving in the learning of Science. There may be another advantage in the development of such an index. The same PP-plots for the same variable can be generated from the data collected for different international studies and allow for meaningful comparisons of the data collected in these studies.

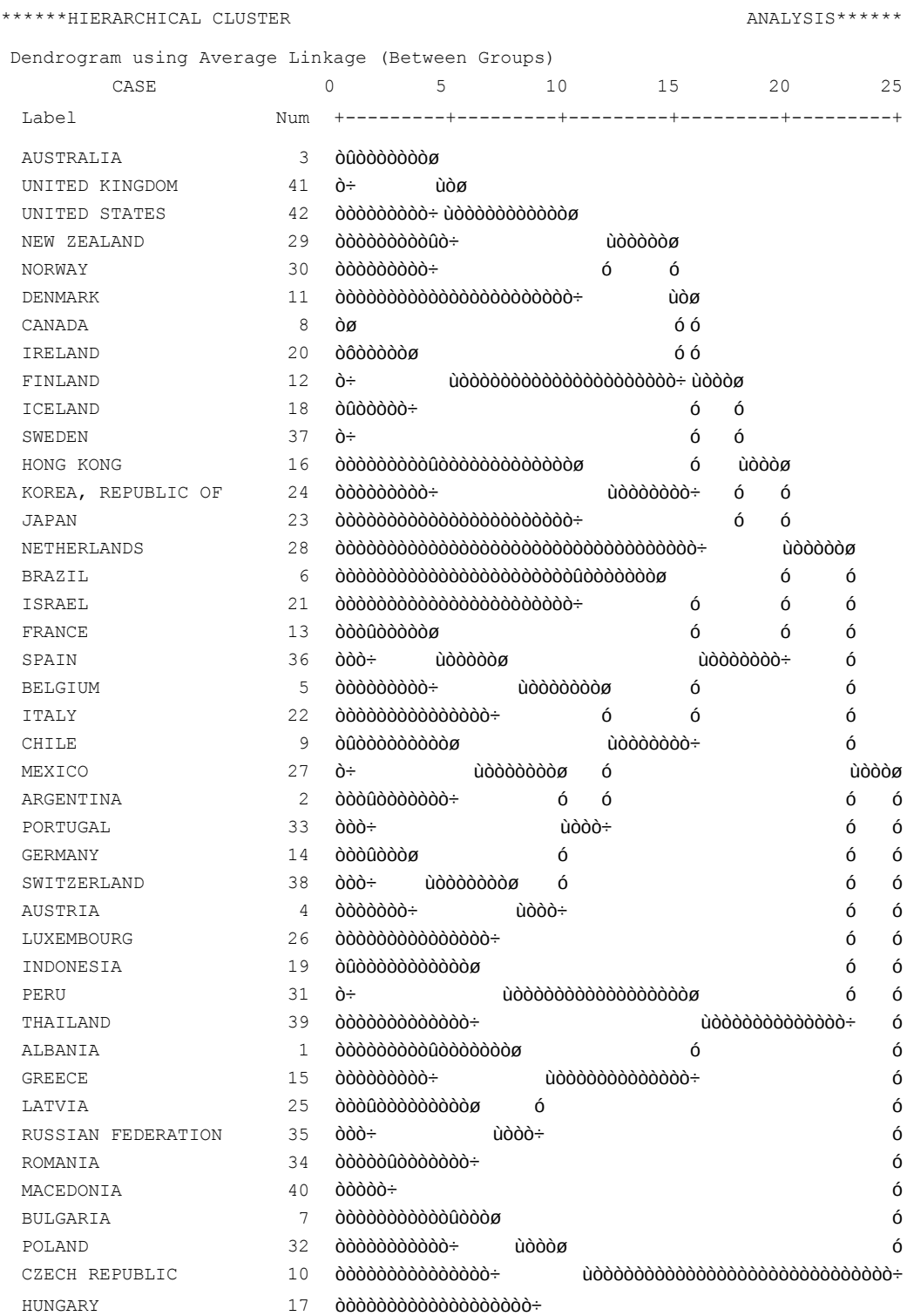


Figure 11 Dendogram generated after Hierarchical Cluster Analysis with unstandardised regressions coefficients for all national samples (File 1.6)

CONCLUSIONS

One question is of great interest. Do the research findings from one country apply to another country? This question is particularly important in the light of limited human and financial resources for educational research.

For example, for three countries from the PISA survey: A, B and C, Science achievement when regressed on variable X may yield similar values of a metric regression coefficient at the whole population level. It would not be enough though, to apply the policy conclusions from research in a field connected with the variable X, that were made in country C, to both countries A and B. However, when examining Figure 12, which present the PP-plot for the whole range of achievement levels, the graphs for countries A and C are very close to each other and both very different from the graph for country B. Would it be more justified to argue, on the base of similar PP-plot shapes, that a particular variable influences Science achievement in a similar way for these two countries? Would it be more legitimate to apply policy conclusions from research in a field connected with the variable X in an exchangeable way between countries A and C? This is a very simplified example, but these unanswered questions are of considerable importance in comparative research in the field of education, especially since Peru and Thailand are both developing countries with limited resources for research in education. However, because of the extensive body of educational research in certain highly developed countries, it is commonly assumed that similar relationships apply in developing countries.

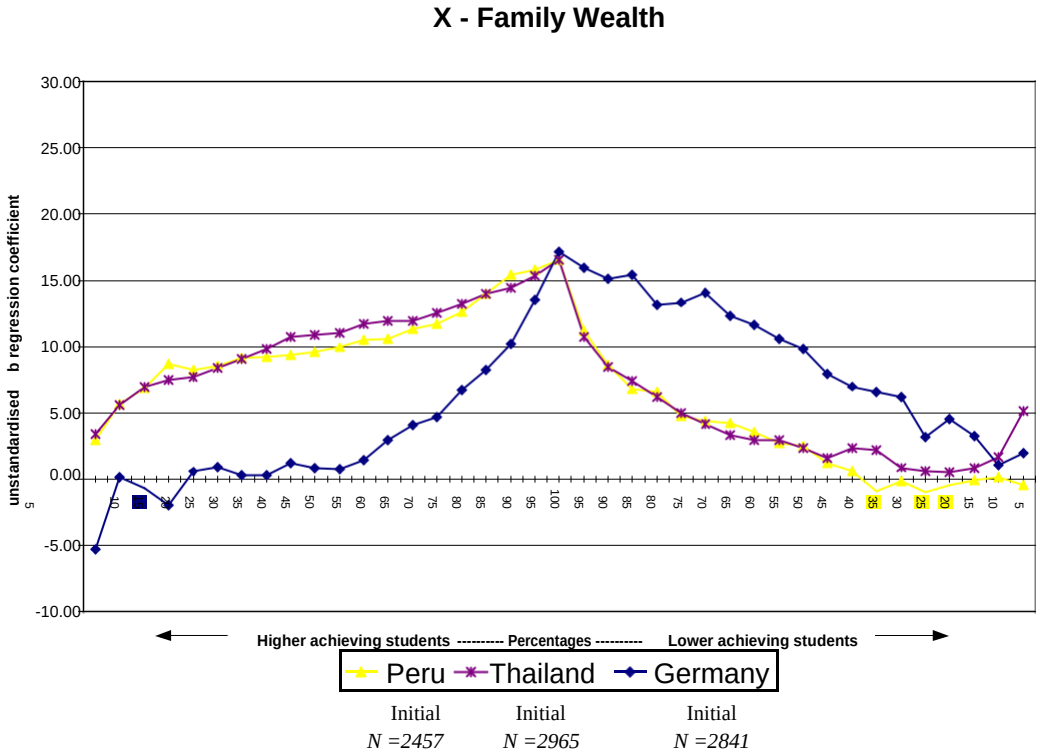
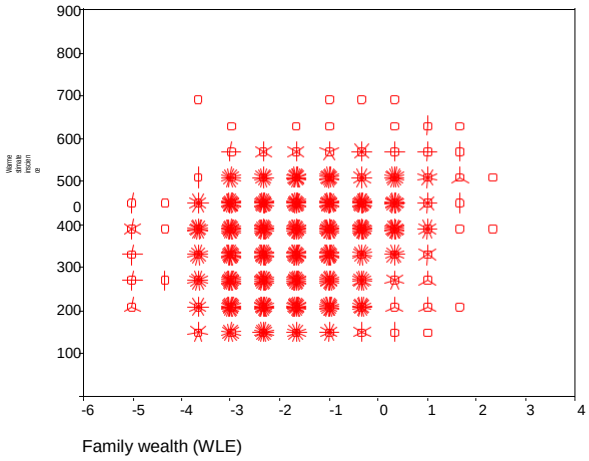


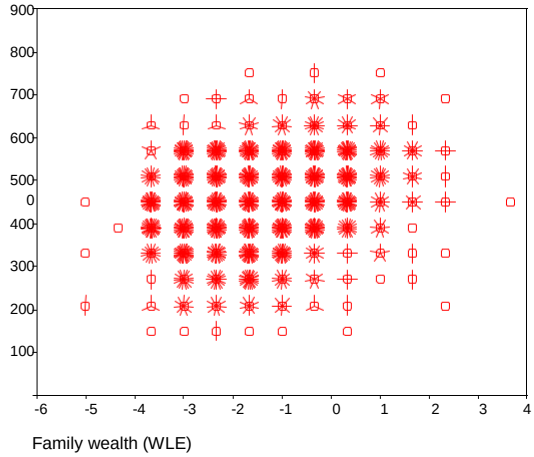
Figure 12 Plot of the unstandardised regression coefficients for Science achievement regressed on the independent variable X – Family wealth for different Science achievement subgroups and A – Peru B - Germany C - Thailand (Source file A1.3)

In the same way as is discussed in the Family wealth section, in addition to the PP-plots in Figure 12 the appropriate scatter plots were generated and are presented in Figure 13. The first two with very similar shapes were generated for Peru and Thailand and the third for Germany. The similarities between scatter plots for Peru and Thailand seem to be obvious and support those observed with PP-plots. However, the decision about the similarity between scatter plots is based on a subjective judgment, when PP-plots permit approaching this problem in a way that leads more readily to calculation.

PERU



THAILAND



GERMANY

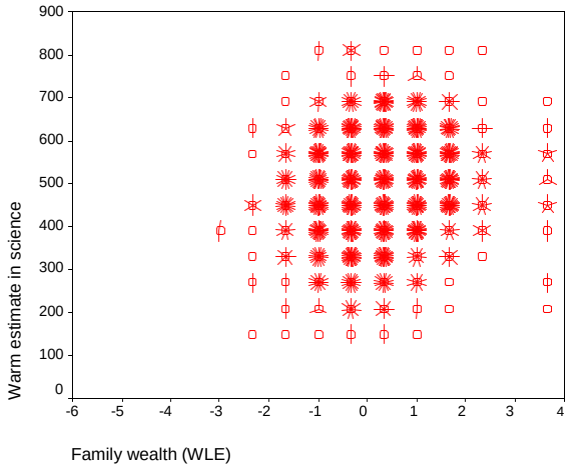


Figure 13 Scatter plots with relationship between Family wealth and Science achievement for Peru, Thailand and Germany respectively.

The main ideas that are at the forefront of a proposed strategy which involves the use of PP-plots can be stated in questions: How can countries be compared in the magnitude to which a particular variable influences Science achievement and how can this comparison be made across all achievement levels and the important subgroups of high and low performing students? In this article a new strategy has been introduced that may be the first step towards obtaining such a two dimensional comparison. It has to be noted, however, that through the PP-plots it is possible to investigate only how one variable at a time influences Science achievement, although it does examine data collected from different countries and for different achievement subgroups. This means that the many analytic possibilities that are available through using multivariate and multilevel analyses are not used here at all.

An interesting and important extension of the idea underlying the formation of fractiles using PP-plots, is to extend this idea to the analyses of simple multivariate and multilevel models that are

tested initially with partial least squares programs which are robust under the conditions of lack of normality in the score distributions. It is highly probable that very different factors operate to influence both the achievement and attitudinal scores in science and mathematics of very high and very low performing students. These issues must be addressed in the cross-national testing programs in addition to the simplistic, although highly accurate estimation and ranking of national mean scores.

REFERENCES

Adams, R., & Wu M. (eds.) (2002). PISA 2000 Technical Report. Paris: OECD.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., et al. (2000). TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade. Chestnut Hill, MA: Boston College.

28th International Physics Olympiad. Sudbury, Ontario, Canada. 1997. [Online]. Available: http://laurentian.ca/physics/OLYMPIAD/PROCEED/RESULTS/INDEX.HTML#gold [2006 , April 11].

XXIX International Physics Olympiad. Taeknigardur, University of Iceland, Reykjavik. 1998. [Online]. Available: http://www.hi.is/pub/ipho/index.html [2006 , April 11].

XXX International Physics Olympiad. Padova, Italy. 1999. [Online]. Available: http://www.pd.infn.it/Olifis/welcome.htm [2006 , April 11].

XXXI International Physics Olympiad. University of Leicester, Leicester. 2000. [Online]. Available: http://www.star.le.ac.uk/IPhO-2000/results_silver.html [2006 , April 11].

XXXII International Physics Olympiad. Antalya, Turkey. 2001. [Online]. Available: http://www.ipho2001.org.tr/results/gold.html [2005 , January 22].

33rd International Physics Olympiad. Bali, Indonesia. 2002. [Online]. Available: http://www.fi.itb.ac.id/~ipho33/results.html [2006 , April 11].