

**TIMSS 2007 secondary analysis: A method
for investigating an attainment gap**

Linda Sturman and Yin Lin

NFER

National Foundation for Educational Research

UK

Paper presented at the 4th IEA International Research Conference (IRC-2010)

University of Gothenburg

1-3 July 2010

Project Team

Project Director

Liz Twist

Project Leader

Linda Sturman

Project statisticians

Yin Lin

Ben Styles

Francesca Saltini

Review

Graham Ruddock



Contents

Abstract

1	Introduction and research question	1
2	Methodology	1
2.1	The variables in the models	1
2.2	Building the models	6
2.3	Presenting and interpreting the outcomes	8
3	Issues and decisions.....	10
3.1	Technical challenges	10
3.2	Interpreting findings	12
4	So, is the effort worth it?.....	13
5	References and further reading.....	14
	Appendix A: Effect size charts	15
	Appendix B: Summary tables of findings	19

Abstract

This paper describes a method of analysis used to explore the TIMSS 2007 achievement gap in eighth grade mathematics between a high-scoring country and the group of Asian Pacific Rim countries which outscored it. A multi-level modelling approach was used to identify key factors significantly associated with attainment in these countries, with the aim of characterising how *high* attainment differs from *the highest* attainment. The analysis used attainment data and background variable data from the TIMSS 2007 international database. Factor analysis was conducted on the background variables and the resulting factors and other relevant background variables were used to create parallel multi-level models. These investigated attainment in the target country (England) and the five comparator countries (Chinese Taipei, Republic of Korea, Singapore, Hong Kong SAR and Japan). This paper outlines the methodology used in the analysis, and shows the outcomes across the six countries. It discusses some of the advantages, challenges and issues associated with this type of analysis, and indicates how the outcomes can be used to attempt to explain an achievement gap, thus helping to inform policy and practice and, potentially, to raise achievement.

Keywords: TIMSS; mathematics; achievement; multi-level modelling; high attainment.

1 Introduction and research question

One of the key findings from TIMSS 2007 for England was that, while its students performed well in mathematics and science at both fourth grade (G4) and eighth grade (G8), they were consistently outperformed by a group of Asian Pacific Rim countries: Chinese Taipei, Republic of Korea, Singapore, Hong Kong SAR, and Japan. The gap between England's achievement and that of these other countries was largest in mathematics at G8 (England's mean in G8 mathematics was 513, significantly higher than the scale average of 500, but significantly lower than the nearest Pacific Rim country, Japan at 570, and the highest scoring country, Chinese Taipei at 598). The gap for mathematics at G4 was smaller (England 541, Japan 568, Hong Kong 607) as were the gaps for science at both grades. This raised the question of why the gap for G8 mathematics might be so large and what could be done to close it.

The international TIMSS reports (Martin *et al*, 2008, Mullis *et al*, 2008) identify variables that vary with attainment. However, England's national multi-level modelling analysis following TIMSS 2007 indicated that some variables which varied with mathematics attainment nevertheless did not show a significant relationship with that attainment once other related factors were controlled for (Sturman *et al*, 2008). The usefulness of the multi-level modelling approach within a country prompted discussion about the possibility of using such an approach across countries, as a means of investigating how *high* attainment might differ from *the highest* attainment. The aim of this analysis was to identify trends in the highest performing countries which might help to explain the attainment gap and, therefore, to inform policy or practice in England and potentially raise achievement.

This paper explores the methodology used in addressing that aim, outlines issues that arose during the analysis, and briefly describes the outcomes.

2 Methodology

2.1 The variables in the models

The first step in this analysis was to specify background variables of interest from the TIMSS 2007 international database (Foy *et al*, 2008). Relevant variables which had proved significant in England's 2008 multi-level modelling analysis were included, as were others hypothesised to be potentially significant in predicting attainment in the Pacific Rim comparator countries. These variables were drawn from responses to the TIMSS student questionnaires, mathematics teacher questionnaires and school questionnaires. Table 1 shows the number of respondents in each country. Although the sample sizes vary somewhat, multi-level modelling is not greatly affected by differences in sample size: strong relationships can still be identified, as can the direction of those relationships.

The variable specification resulted in a large number of variables of interest. These needed to be reduced in number, which was achieved through factor analysis. This technique groups variables (in this case, responses to questions) that behave ('load') in similar ways. These groups form factors which are entered into the models, reducing the variables to a manageable number.

Table 1: Sample sizes

Country	Number of:		
	Students	Maths teachers	Headteachers
Chinese Taipei	4046	152	150
Korea	4240	243	150
Singapore	4599	357	164
Hong Kong SAR	3470	145	120
Japan	4312	216	146
England	4025	235	137

A key element in this investigation was to identify factors that loaded in the same way across the target countries, so that robust comparisons could be made across them. Therefore, exploratory factor analysis was initially applied to responses from the school questionnaire (HQ), mathematics teacher questionnaire (TQM) and student questionnaire (SQ). Factor analysis was conducted for England and each of the five Pacific Rim countries separately and similarities and differences in their question response patterns were observed. Where question response patterns were similar, common factors across the six countries were constructed using the same questions or items, allowing direct comparisons in further analysis. The group of related questions or items in each factor was then scaled to give a factor score from 0 to 10. Scaling was used so that it was possible to compare each factor's mean score with that of other factors, hence evaluating the relative strength of responses.

The small number of target variables which did not load were reviewed and either excluded from the models or, in most cases, refined and entered into the model as revised factors or separate variables. The resulting factors and other relevant background variables were then used to create parallel multi-level models.

Using the factors and the separate variables gave rise to a large number of possible predictors, some of which were highly correlated. When highly correlated predictors are entered into a model they can interfere with each other ('multicollinearity'), to produce false results. In order to avoid multicollinearity, a statistical measure called 'tolerance' was employed to identify any problematic predictors. Tolerance is a measure of the proportion of variance in a predictor which cannot be explained by other predictors in the model. For any predictor, a low tolerance value suggests that a large proportion of variance in this predictor can be explained by other predictors in

the model, indicating that multicollinearity is present. Therefore, predictors with low tolerance should be excluded. In this study, problematic predictors were removed one by one, until all remaining predictors had tolerance above 0.2 (and mostly above 0.3) across all six countries.

The selected variables and the factors are given in Tables 2.1 to 4.2 below.

Table 2.1 Student questionnaire (SQ) factors

Scale label	Questionnaire items	No. of items in factor
SQ9 - Valuing maths	SQ9 (a, b, c, d)	4
SQ15 - School climate	SQ15 (a, b, c)	3
SQ17 - Out of school activities - socialise/technology	SQ17 (a, b, c, h)	4
SQ8 - Confidence and enjoyment in maths	SQ8 (a, b, c, d, e, f, g, h)	8
SQ16 - During last month - incidents at school	SQ16 (a, b, c, d, e)	5

Table 2.2 Student questionnaire (SQ) variables

Scale label	Questionnaire item
Gender of student – boy	itsex
SQ4 - Books in home	SQ4
SQ5c - Resources at home - Study desk	SQ5c
SQ5e - Resources at home - Internet	SQ5e
SQ7 - Educational aspiration	SQ7
SQ10k - Lecture-style presentation	SQ10k
SQ10l - Independent working	SQ10l
SQ17g - Out of school - read book for enjoyment	SQ17g
SQ18a - Frequency of maths homework	SQ18a
SQ18b - Time on maths homework	SQ18b
SQ21 - Student's time in country	SQ21

Table 3.1 Mathematics teacher questionnaire (TQM) factors

Scale label	Questionnaire items	No. of items
TQM7 - Preparedness to teach maths topics	TQM7 (all items)	18
TQM8 - Teacher interaction - discussion and preparation	TQM8 (a, b)	2
TQM8 - Teacher interaction - observation	TQM8 (c, d)	2
TQM10 - Safety at school	TQM10 (a, b, c)	3
TQM11 - Adequate buildings and space	TQM11 (a, b, c)	3
TQM17 - Maths activities – domains	TQM17 (b, c, d, e)	4
TQM17 - Maths activities – methods	TQM17 (f, g, h, i, j, k)	6
TQM18 - Fewer limitations - students	TQM18 (a, b, c, d, e)	5
TQM18 - Fewer limitations - resources	TQM18 (f, g, h, i, j, k, l, m)	8
TQM22 - Calculator usage – routine	TQM22 (a, b)	2
TQM22 - Calculator usage – complex	TQM22 (c, d)	2
TQM20 - Topic coverage	TQM20 (all items)	39
TQM9 – CPD	TQM9 (a, b, c, d, e, f)	6

Table 3.2 Mathematics teacher questionnaire (TQM) variables

Scale label	Questionnaire item
Gender of maths teacher – male	TQM2
TQM3 - Teaching experience	TQM3
TQM5 - Teacher specialism - maths education (<i>in comparison to 'teacher specialism - mathematics'</i>)	TQM5
TQM5 - Teacher specialism - others (<i>in comparison to 'teacher specialism - mathematics'</i>)	TQM5
TQM15 - Textbook use - supplementary (<i>in comparison to 'Textbook use - primary basis'</i>)	TQM15
TQM15 - Textbook use - none (<i>in comparison to 'Textbook use - primary basis'</i>)	TQM15
TQM16g - Percentage of class time - off task	TQM16g

Scale label	Questionnaire item
TQM17l - Maths activities - work in small groups	TQM17l
TQM21 - Calculator - less restricted use	TQM21
TQM26 - Frequency of maths homework	TQM26
TQM27 - Time required for maths homework	TQM27
TQM28a - Homework type - question sets	TQM28a
TQM28b - Homework type - gather data and report	TQM28b
TQM28c - Homework type - finding applications	TQM28c
TQM29b - Homework use - correct and give feedback	TQM29b
TQM29d - Homework use - class discussion	TQM29d
TQM30a - Monitoring progress - Classroom tests	TQM30a
TQM30b - Monitoring progress - National/regional tests	TQM30b
TQM30c - Monitoring progress - Professional judgement	TQM30c
TQM31 - Frequency of maths tests	TQM31
TQM32 - Test items - multiple choice	TQM32
TQM33a - Test items – recall	TQM33a
TQM33d - Test items - explanation and justification	TQM33d

Table 4.1 School questionnaire (HQ) factors

Scale label	Questionnaire items	No. of items
HQ18A - Frequency of minor problems	HQ18A (a, b, c, d)	4
HQ18A - Frequency of serious problems	HQ18A (i, j, k, l)	4
HQ8 - School climate - teachers and parents	HQ18A (a, b, c, d, i, j, k, l)	8

Table 4.2 School questionnaire (HQ) variables

Scale label	Questionnaire item
HQ3a - School roll - fewer economically disadvantaged students	HQ3a
HQ3b - School roll - more economically affluent students	HQ3b
HQ5 - School hours per week	HQ5 (b, c)
HQ9 - Maths ability groups	HQ9
HQ10a - School offers enrichment maths	HQ10a
HQ10b - School offers remedial maths	HQ10b
HQ14a - Maths teacher evaluation - Internal observations	HQ14a
HQ14b - Maths teacher evaluation - External observations	HQ14b
HQ14c - Maths teacher evaluation - Student achievement	HQ14c
HQ14d - Maths teacher evaluation - Peer review	HQ14d
HQ16a - Difficulty in filling maths teacher vacancies	HQ16a
HQ17a - Incentives to recruit or retain maths teachers	HQ17a

2.2 Building the models

Once factors and variables had been finalised, TIMSS 2007 G8 mathematics achievement data for each country was sourced from the international database for building the models. The achievement data consisted of five plausible values (estimated scale scores, see Olson *et al*, 2008 for more information) for each student. Sensitivity analysis in an earlier study (see Sturman *et al*, 2008, technical appendix) confirmed that taking the mean of these five values as the outcome variable in the model would lead to underestimation of standard errors. In light of this, a multi-level model was built for each of the five plausible values and the final model coefficients were obtained by averaging the results of these five models. In addition, so as to get final estimates of standard errors for the model coefficients, imputation error estimates calculated from the five models were combined with sampling error estimates calculated using jackknifing techniques.

Finally, because the number of target variables was large, a two-stage process of modelling was used. At the first, interim, stage the same variables were entered for all six countries and a ‘sifting’ process, using a generous significance level (10 per cent), identified the variables potentially significant for each country.

In the interim models, the generous significance level was used so that potentially significant variables could be separated from those unlikely to be statistically

significant. This enabled a smaller number of potentially key variables to be entered into the final models, where the standard significance level of 5 per cent was used.

The two-stage process was carried out as follows:

1. Interim model for each country:
 - a. Run multi-level models for each of the five plausible values, using all selected predictors;
 - b. Average coefficients from these five models to get the interim model;
 - c. Calculate imputation error using results from these five models;
 - d. Apply jackknifing to the model with the first plausible value to estimate sampling error;
 - e. Combine imputation error and sampling error to get standard error estimates for the interim model;
 - f. Identify predictors which were significant at the 10 per cent level.
2. Final model for each country:
 - a. Run multi-level models for each of the five plausible values, using only the significant predictors identified in the interim model;
 - b. Average coefficients from these five models to get the final model;
 - c. Calculate imputation error using results from these five models;
 - d. Apply jackknifing to the model with the first plausible value to calculate sampling error;
 - e. Combine imputation error and sampling error to get standard error estimates for the interim model;
 - f. Identify predictors which were significant at the 5 per cent level.

For each of the six countries, although the same predictors were fed into the interim models, different predictors could be identified as significant for each country's final model. Thus, the factors and variables entered into the final model for each country could differ. Nevertheless, comparisons can still be made between the final models as only the significant variables have any strong association with the outcomes. Thus, the inclusion of variables identified as non-significant in any given interim model would not have added any improvement to the results if entered into the final model for that country.

In order to explore variations at each level explained by the predictors and to enable calculation of quasi-effect size coefficients later, 'base case' models with no predictors were also constructed.

Multi-level modelling takes account of data which is grouped into similar clusters at different levels. For example, individual students are grouped into classes, and those classes are grouped within schools. There may be more commonality between students within the same class than with students in other classes, and there may be elements of similarity between different classes in the same school. Multi-level modelling takes account of this hierarchical structure of the data and produces more accurate predictions than simple regression analysis. Moreover, it allows the variance in attainment at each level to be estimated.

In this study, the models incorporated two or three levels, depending on the data structure of the particular country. If multiple classes were sampled within each school, the model incorporated three levels: school, class and student. However, if only one class was sampled in most of the schools, the model incorporated just two levels: school and student. The hierarchical structures of the models are detailed in Table 5 below.

Table 5: Hierarchical structures of the models

Country/region	Total schools in the sample	Schools with one class per school in the sample	Hierarchical structure in the model
Chinese Taipei	150	147	School/Student
Korea	150	150	School/Student
Singapore	164	2	School/Class/Student
Hong Kong SAR	120	120	School/Student
Japan	146	123	School/Student
England	137	68	School/Class/Student

2.3 Presenting and interpreting the outcomes

Although variables in the models were controlled for, multi-level modelling was not able to control for other variables which were not in the models. Past analysis in England had shown that prior attainment was a major predictor for attainment and hence any model of attainment should ideally have prior attainment included as a predictor. However, in the case of this study, prior attainment data of students was not available for all countries and hence the models were not able to control for its influence. As a result, care must be taken when interpreting the relationships, especially with predictors which might link strongly with prior attainment.

To aid interpretation, the model results had been converted into ‘quasi-effect size coefficients’ (Schagen and Elliot, 2004). Quasi-effect size coefficients represent the expected changes (in percentage of the standard deviation) in the outcome for an average switch between low and high values in the predictor variables. In other words, if a predictor has a relatively large quasi-effect size coefficient (positive or negative), then change in this predictor will be associated with a relatively large change in mathematics attainment across the population as a whole.

Quasi-effect size coefficients are plotted in Appendix A for the final model of each country. Only significant predictors (at the 5 per cent level) are plotted. For each predictor, the estimated quasi-effect size coefficient is plotted as a diamond, with a vertical line indicating the 95 per cent confidence interval for the estimate. Positive values imply a positive relationship with the international scale outcome; negative values imply that the outcome scale values tend to decrease with higher values of the given predictor variable.

It should be emphasised that the results of these models document associations and not causal relationships. Also, in any model, it is expected that some variables would come out significant by chance. Thus, it is possible that some of the borderline significant effects in these models might not be genuine.

Appendix B summarises the findings across the six countries. The comparative tables shown in Appendix B are being used as the basis for exploring how the outcomes might explain the achievement gap between England and the Pacific Rim countries outscoring England in mathematics at G8. The areas of key interest are those in which relatively strong effects, in the same direction, were found for several countries. These include: G8 students’ levels of confidence and enjoyment in mathematics; their educational aspirations; the number of books they have in the home; the school climate and incidents at school (i.e. the learning and social environment at their schools); their teachers’ reports of limitations on their teaching caused by the range of students in their class; and the extent of previous topic coverage as reported by mathematics teachers. These are summarised in Table 6 below, with their effect sizes and directions. More information is given in Appendix B.

Other findings are being explored too. These include considering whether effects found in only one or two countries might hold some lessons for policy or practice, or whether effects which vary in direction might also hold such lessons. In addition, the potential relevance of the absence of an effect for particular variables, for which effects might have been anticipated, is being considered.

Table 6: Particularly strong and/or consistent effects in the models

Variable	Quasi-effect size:					
	England	Chinese Taipei	Korea	Singapore	Hong Kong SAR	Japan
SQ School climate	-3	-8	-7		-3	-5
SQ Confidence and enjoyment in maths	+17	+35	+37	+25	+25	+33
SQ During last month - incidents at school	-4	-2			-4	-5
SQ Books in home	+9	+12	+15			+11
SQ Resources at home - Internet		+7	+9	+2	+2	+9
SQ Educational aspiration	+5	+19	+11		+7	+23
SQ Independent working	+2	+10	+21		+3	+13
TQM Topic coverage	+34	+6		+21	+13	+5
TQM Fewer limitations - students	+19		+5	+14	+13	

A positive relationship indicates that the variable is associated with higher attainment; a negative relationship that it is associated with lower attainment.

Moderate or large quasi-effect sizes (5 points upwards) are more likely to indicate real effects; some smaller values might indicate borderline or spurious relationships, which can arise in any model.

3 Issues and decisions

3.1 Technical challenges

Building parallel multi-level models across six countries raises a few technical challenges. Firstly, decisions are needed about which variables to investigate. Which variables to consider will obviously depend on the research question, and what data is available for each country. Taking this into account, the more information the analysis team has about the context in each country under investigation, the easier it becomes to predict which variables would be worth investigating across the countries and which can safely be excluded. In other words, there needs to be some hypothesis about which variables might show effects and which might not.

The next decision, when conducting factor analysis, is how to deal with questions and items that load differently across the countries. Unless countries are remarkably similar, there will always be different patterns of response to some questions. In the current study, when loading of a group of items showed some consistency but not enough to form identical factors across countries, several approaches were adopted to address this. The approach selected in each case depended on the circumstances. In some cases, the items loading inconsistently were reviewed and some less relevant items were removed to see if consistency and reliability improved. Alternatively, one or more items were extracted from the group and used as separate variables in further

analysis (subject to not being highly correlated with each other or the factor). Another alternative was to ‘force’ such items into parallel factors, subject to the response patterns being close enough to accommodate such action, and the reliability figures for the resulting factors being sufficiently high to allow confidence in the measure. As a last resort, an item would be omitted from further analysis. This was rarely done in the case of this study: a suitable compromise could generally be found for variables felt to be of key importance.

Once factors and separate variables are decided upon, the potential issue of multicollinearity needs addressing. Highly correlated variables can interfere with each other and produce false multi-level modelling results. This is particularly likely in models with a large number of variables. In the case of this study, high correlations were a potential issue for questionnaire items that did not load neatly onto factors and might be entered into the model separately. They were also a potential issue for multiple sub-items from a single questionnaire item that might be considered as separate variables. In such situations, it is important to check correlations and tolerances for all variables and, where appropriate, revise the variable list accordingly.

Just as parallel factors need to be constructed in all countries, correlations and tolerances also need to support the creation of a robust model for each country. If a variable has particularly low tolerance in one country, it needs to be changed in all countries, even if it is working well in most countries.

In technically challenging situations like those described above, care is needed to ensure that technical issues are resolved in ways that maximise statistical integrity whilst also addressing the research questions. In other words, technical decisions taken to improve the model should not undermine the models’ ability to address the research questions. Statistical integrity and research quality need to remain in balance. Depending on the nature of the challenges, achieving this can be time-consuming.

Once these hurdles have been overcome, the multi-level models can be prepared. However, using scaled achievement data containing plausible values has further time implications. As noted earlier, for each model of each country, parallel models need to be run on each plausible value (in this case, five). The results then need to be averaged in order to obtain appropriate coefficient estimates. Moreover, jackknifing procedures are needed in order to estimate sampling error. This requires that each model, using the first plausible value, should be run many times with different weights (in this case, between 60 to 75 times for each model). This process is time-consuming but necessary for accuracy. Technical information on working with plausible values can be found in the TIMSS Technical Report (Olson *et al*, 2008) and the international database User Guide (Foy *et al*, 2008).

In an analysis of this kind, groundwork is crucial. Giving time to planning and exploratory analysis is vital if the final models are to be robust and useful.

3.2 Interpreting findings

The outcomes from any model are determined by the inputs, which are in turn determined by the purpose of the investigation. An issue that arose in this study was that multi-level models had already been run for England's national report for TIMSS 2007, using variables drawn from the international database and also nationally available data, such as students' prior attainment. The purpose of one of those models was to explore relationships between key background variables and attainment in mathematics at G8. However, since the purpose of this current study was to compare multi-level modelling outcomes across six countries in which comparable national data was not always available, England's model needed to be re-run using comparable inputs. Inevitably, the results of England's two sets of multi-level modelling results were not identical. Differences in results can potentially raise questions, especially where outcomes are markedly different, and any reporting of the results might need to anticipate and address such questions. Reassuringly, in this case, where comparable variables were included in both the original and current G8 mathematics attainment model in England, outcomes were sufficiently similar to require little explanation of that type.

Even so, other challenges arise in interpretation. Causality cannot generally be assigned to outcomes, however tempting it might be to do so. The outcomes describe associations, not causal relationships. For example, where attainment in a subject is positively associated with confidence in learning that subject, the results cannot say whether confidence leads to higher attainment, higher attainment leads to confidence or whether a third variable causes both. Hypotheses about causality can be presented based on the outcomes, but definitive statements about causality cannot be made. Care is therefore needed in presenting results.

Furthermore, interpretation in a model of the kind used in this study is complex. A comparative analysis of this type is unlikely to show neat findings, particularly across six countries. In this analysis, findings were far from straightforward to interpret. Few variables showed consistent effects across all countries, or across the five comparator countries compared with England. In some cases, effects seen in several countries were in opposite directions and/or the strength of the effects varied. Thus, in such situations, it is not easy to draw out messages that might support developments in policy or practice.

In addition, describing and attempting to explain outcomes from parallel multi-level models requires reference to more sources of information than just the model outcomes themselves. The frequencies for responses to the questions underlying each model can help to build up a picture of the context that supports interpretation of the finding. Measures of correlation and tolerance might also need to be reviewed, particularly if findings are unexpected (e.g. effects for a single variable act in different directions in different countries, or a finding seems counter-intuitive, or non-credible). Outcomes should not be taken at face value; the potential for findings to be borderline, or for interference between variables to have arisen must always be borne

in mind. Finally, of course, the different contexts of the countries involved in the analysis need to be taken into consideration when interpreting the results. Different results between two countries might be explained by differences in their education systems. Equally, similar results might have different underlying causes, depending on context.

4 So, is the effort worth it?

Yes!

Although this type of analysis can pose challenges, it can potentially be more informative than simple regression analysis or comparison of frequencies. Its main strengths lie in: being able to investigate multiple relationships between variables and the outcome simultaneously, thus separating out effects; and in being able to quantify variance in the outcome at the different levels of analysis.

This type of analysis is, therefore, worth undertaking, although thought needs to be given to the purpose of the research, the amount of preparation and set-up required, the potential challenges and ways in which these can be met, and issues of interpretation. Our conclusion is that the analysis is feasible, once decisions are made about the approach, the relevant variables, and the construction of the models. The robustness of the decisions made needs to be evaluated at all stages of the process.

The technical decisions involved require familiarity with the international database and the analysis methods used to produce it. However, the technical report (Olson *et al*, 2008) and the User Guide (Foy *et al*, 2008) can support the process.

The resulting outcomes from any models developed might not necessarily answer all research questions. However, they are likely to answer some and will undoubtedly prompt further questions that might help in hypothesising about the research questions, and potentially inform policy and practice to improve students' learning - which is, after all, one of the main aims of international surveys.

5 References and further reading

Foy, P. & Olson, J.F. (Eds.). (2008). *TIMSS 2007 International Database and User Guide*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

IEA International Database Analyzer (IEA IDB Analyzer),
http://www.iea.nl/iea_studies_datasets.html

Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (Published December 2008, revised August 2009)

Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. (Published December 2008, revised August 2009)

Olson, J.F., Martin, M.O., & Mullis, I.V.S. (Eds.). (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Schagen, I. and Elliot, K. (2004) *But what does it mean? The use of effect sizes in educational research*. Slough: NFER.

Sturman, L., Ruddock, G., Burge, B., Styles, B., Lin, Y. and Vappula, H. (2008). *England's Achievement in TIMSS 2007: National Report for England*. Slough: NFER.)

Appendix A: Effect size charts

This appendix presents a series of charts showing the quasi-effect size coefficients for the final model of each country. Only significant predictors (at the 5 per cent level) are plotted. For each predictor, the estimated quasi-effect size coefficient is plotted as a diamond, with a vertical line indicating the 95 per cent confidence interval for the estimate. Positive values imply a positive relationship with the international scale outcome; negative values imply that the outcome scale values tend to decrease with higher values of the given predictor variable.

The outcomes describe associations, not causal relationships. For example, where attainment in mathematics is positively associated with confidence and enjoyment in learning mathematics, the results cannot say whether confidence and enjoyment lead to high attainment, enjoyment and confidence result from high attainment, or whether a third variable causes both.

Figure A1: Final model - Chinese Taipei

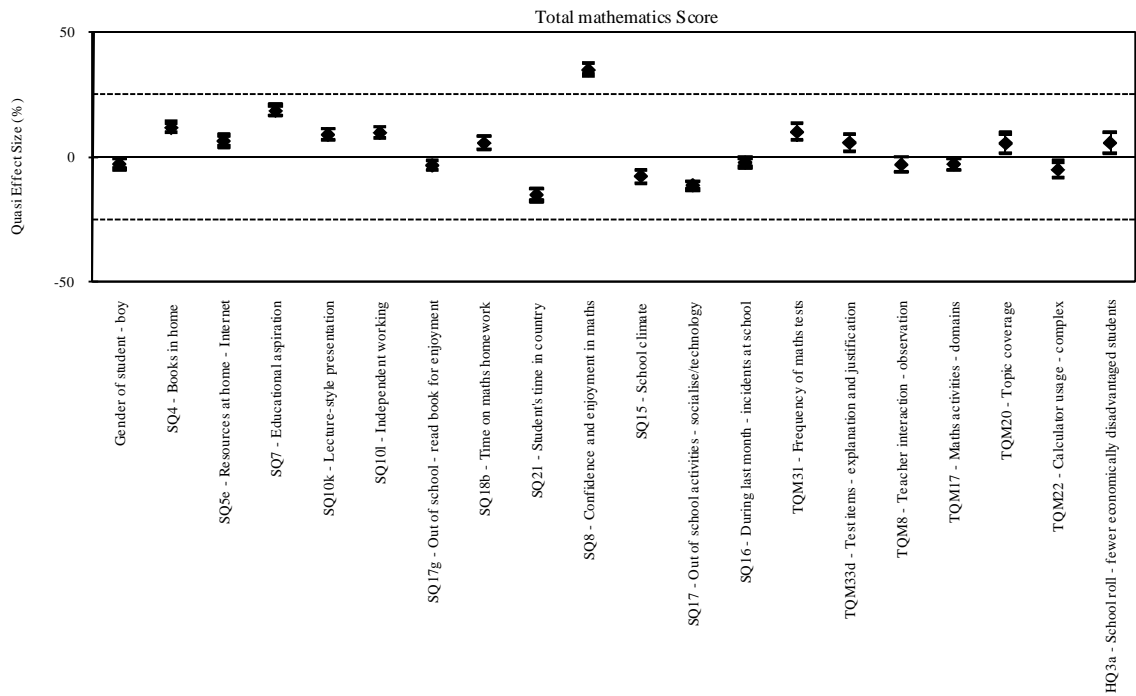


Figure A2: Final model – Korea

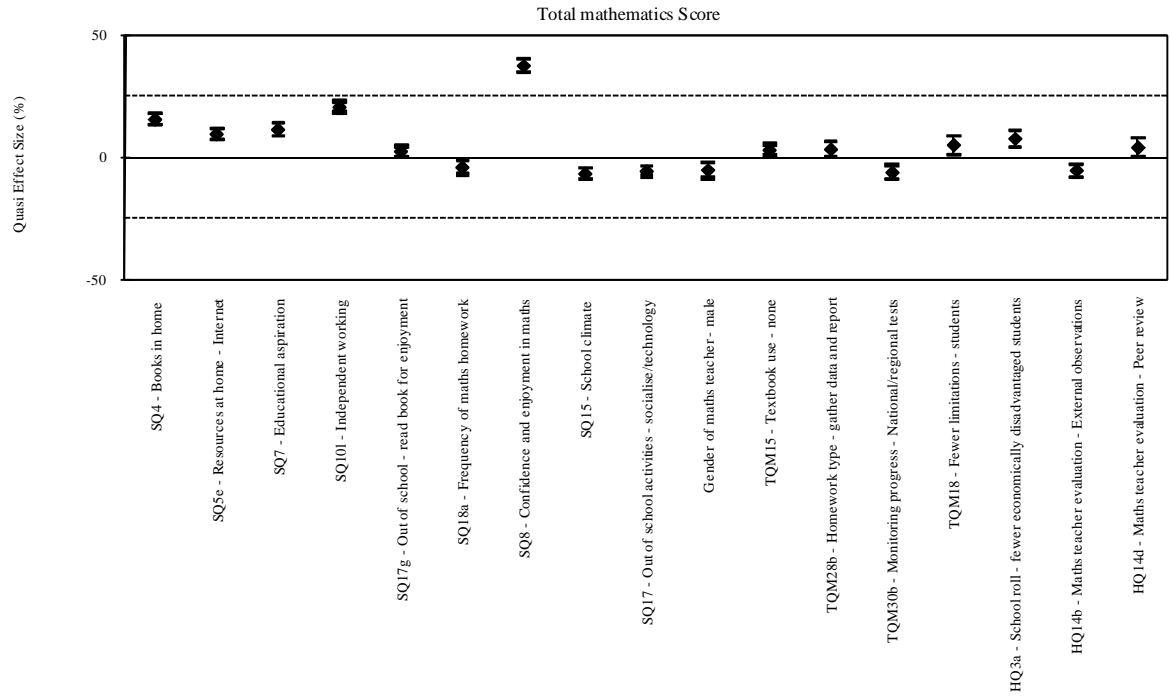


Figure A3: Final model - Singapore

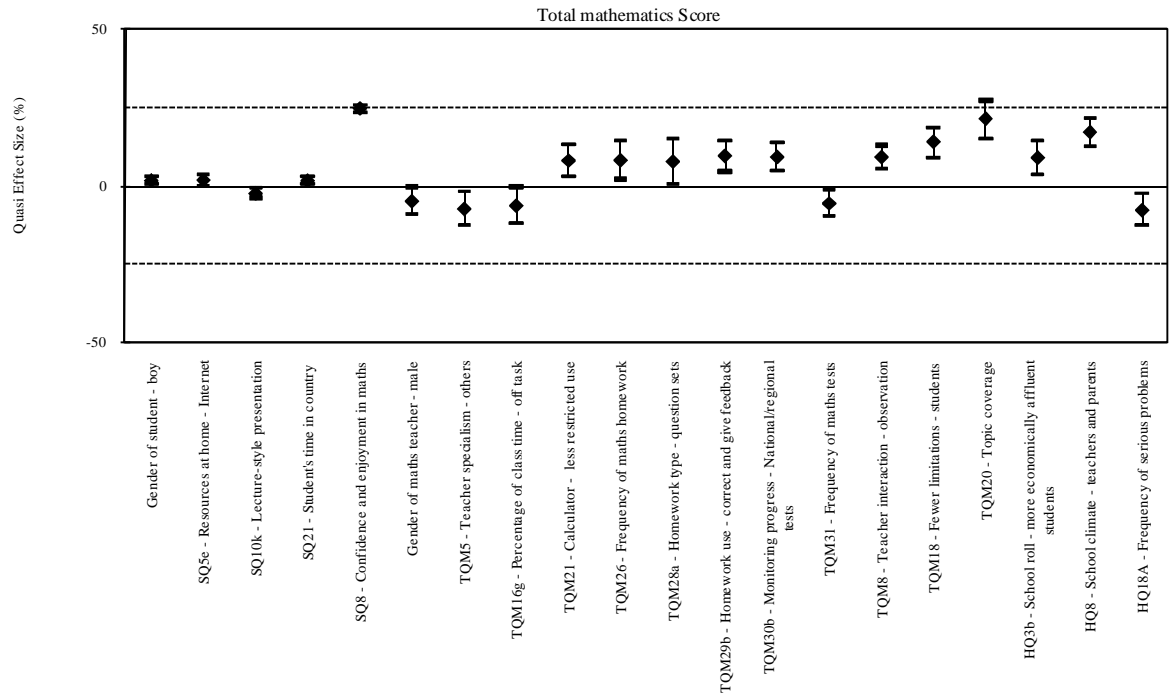


Figure A4: Final model - Hong Kong SAR

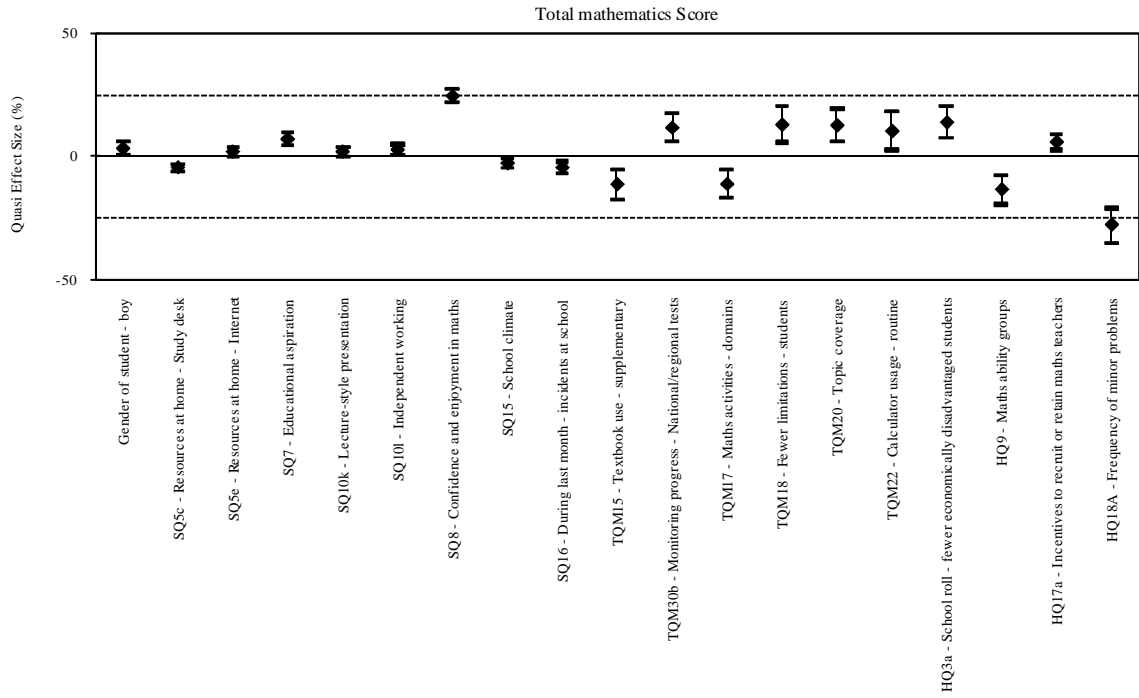


Figure A5: Final model – Japan

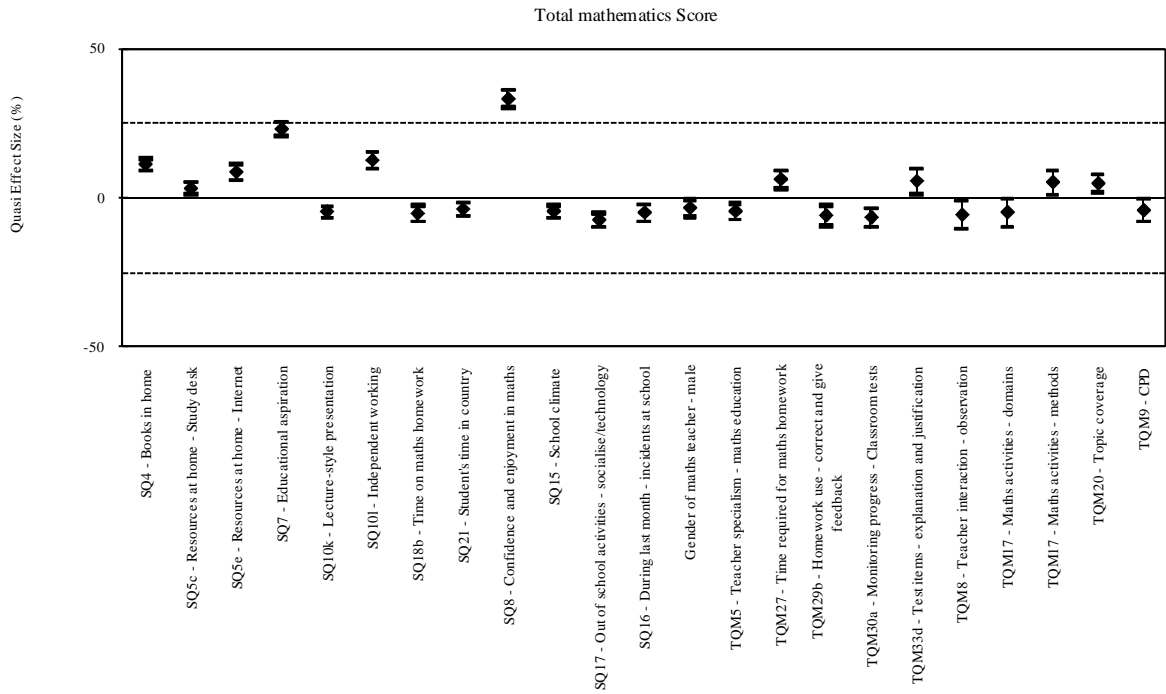
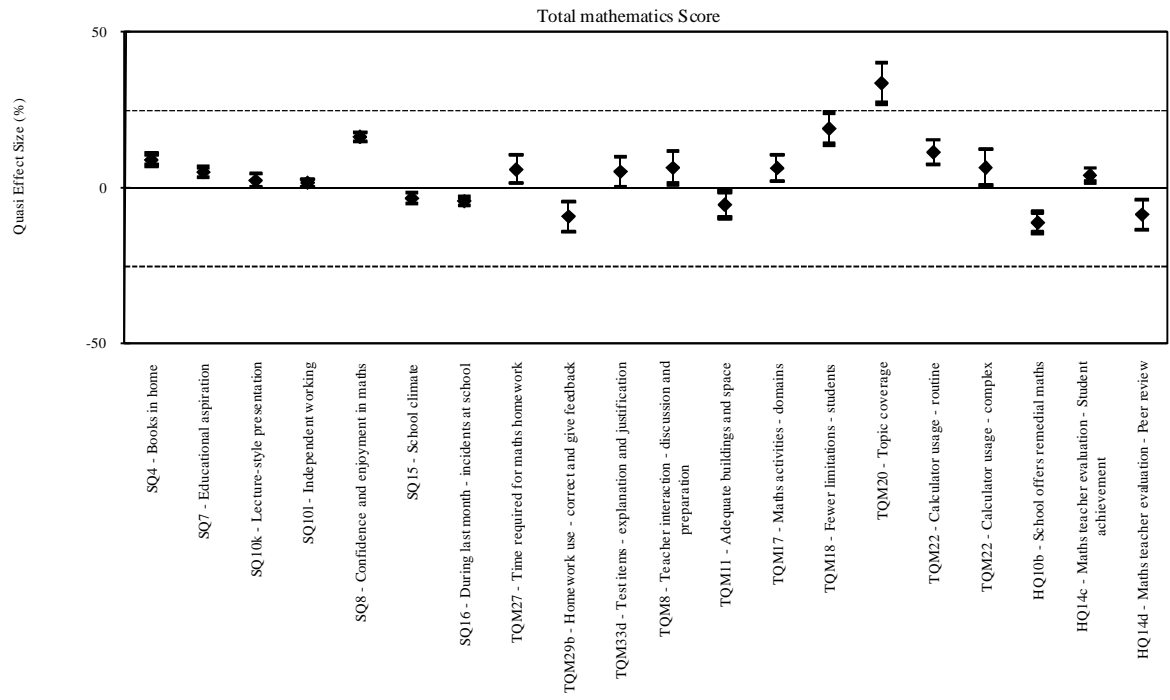


Figure A6: Final model - England



Appendix B: Summary tables of findings

Table B1: Relationships between student attainment and variables derived from the student questionnaire

Student variable (questionnaire factors and items)	Quasi-effect size:					
	England	Chinese Taipei	Korea	Singapore	Hong Kong SAR	Japan
SQ9 (abcd) - Valuing maths						
SQ15 (abc) - School climate	-3	-8	-7		-3	-5
SQ17 (abch) - Out of school activities - socialise/technology		-11	-6			-7
SQ8 (abcdefgh) - Confidence and enjoyment in maths	+17	+35	+37	+25	+25	+33
SQ16 (abcde) - During last month - incidents at school	-4	-2			-4	-5
Gender of student – boy		-3		+2	+3	
SQ4 - Books in home	+9	+12	+15			+11
SQ5c - Resources at home - Study desk					-4	+3
SQ5e - Resources at home - Internet		+7	+9	+2	+2	+9
SQ7 - Educational aspiration	+5	+19	+11		+7	+23
SQ10k - Lecture-style presentation	+2	+9		-2	+2	-5
SQ10l - Independent working	+2	+10	+21		+3	+13
SQ17g - Out of school - read book for enjoyment		-3	+2			
SQ18a - Frequency of maths homework			-4			
SQ18b - Time on maths homework		+6				-5
SQ21 - Student's time in country		-15		+2		-4

A positive relationship indicates that the variable is associated with higher attainment; a negative relationship that it is associated with lower attainment.

Shaded row(s) indicate a variable that was not statistically significant in any model.

Factors are presented first (questionnaire items included in each factor are indicated) followed by separate variables entered into the model.

Moderate or large quasi-effect sizes (5 points upwards) are more likely to indicate real effects; some smaller values might indicate borderline or spurious relationships, which can arise in any model.

Table B2: Relationships between student attainment and variables derived from the mathematics teacher questionnaire

Teacher variable (questionnaire factors and items)	Quasi-effect size:					
	England	Chinese Taipei	Korea	Singapore	Hong Kong SAR	Japan
TQM7 (all items) - Preparedness to teach maths topics						
TQM8 (ab) - Teacher interaction - discussion and preparation	+7					
TQM8 (cd) - Teacher interaction - observation		-3		+9		-6
TQM10 (abc) - Safety at school						
TQM11 (abc) - Adequate buildings and space	-5					
TQM17 (bcde) - Maths activities - domains	+7	-3			-11	-5
TQM17 (fghijk) - Maths activities - methods						+5
TQM18 (abcde)- Fewer limitations - students	+19		+5	+14	+13	
TQM18 (fghijk) - Fewer limitations - resources						
TQM22 (ab) - Calculator usage - routine	+12				+10	
TQM22 (cd) - Calculator usage - complex	+7	-5				
TQM20 (all items) - Topic coverage	+34	+6		+21	+13	+5
TQM9 (abcdef) – CPD						-4
Gender of maths teacher - male			-5	-5		-3
TQM3 - Teaching experience						
TQM5 - Teacher specialism - maths education (in comparison to ‘teacher specialism - mathematics’)						-5
TQM5 - Teacher specialism - others (in comparison to ‘teacher specialism - mathematics’)				-7		
TQM15 - Textbook use - supplementary (in comparison to ‘Textbook use - primary basis’)					-11	
TQM15 - Textbook use - none (in comparison to ‘Textbook use - primary basis’)			+3			

Teacher variable (questionnaire factors and items)	Quasi-effect size:					
	England	Chinese Taipei	Korea	Singapore	Hong Kong SAR	Japan
TQM16g - Percentage of class time - off task				-6		
TQM17l - Maths activities - work in small groups						
TQM21 - Calculator - less restricted use				+8		
TQM26 - Frequency of maths homework				+8		
TQM27 - Time required for maths homework	+6					+6
TQM28a - Homework type - question sets				+8		
TQM28b - Homework type - gather data and report			+3			
TQM28c - Homework type - finding applications						
TQM29b - Homework use - correct and give feedback	-9			+10		-6
TQM29d - Homework use - class discussion						
TQM30a - Monitoring progress - Classroom tests						-7
TQM30b - Monitoring progress - National/regional tests			-6	+9	+12	
TQM30c - Monitoring progress - Professional judgement						
TQM31 - Frequency of maths tests		+10		-5		
TQM32 - Test items - multiple choice						
TQM33a - Test items – recall						
TQM33d - Test items - explanation and justification	+5	+6				+6

A positive relationship indicates that the variable is associated with higher attainment; a negative relationship that it is associated with lower attainment.

Shaded row(s) indicate a variable that was not statistically significant in any model.

Factors are presented first (questionnaire items included in each factor are indicated) followed by separate variables entered into the model.

Moderate or large quasi-effect sizes (5 points upwards) are more likely to indicate real effects; some smaller values might indicate borderline or spurious relationships, which can arise in any model.

Table B3: Relationships between student attainment and variables derived from the school questionnaire

School variable (questionnaire factors and items)	Quasi-effect size:					
	England	Chinese Taipei	Korea	Singapore	Hong Kong SAR	Japan
HQ18A (abcd) - Frequency of minor problems					-28	
HQ18A (ijkl) - Frequency of serious problems				-8		
HQ8 (abcdijkl)- School climate - teachers and parents				+17		
HQ3a - School roll - fewer economically disadvantaged students		+6	+8		+14	
HQ3b - School roll - more economically affluent students				+9		
HQ5 (bc) - School hours per week						
HQ9 - Maths ability groups					-14	
HQ10a - School offers enrichment maths						
HQ10b - School offers remedial maths	-11					
HQ14a - Maths teacher evaluation - Internal observations						
HQ14b - Maths teacher evaluation - External observations			-5			
HQ14c - Maths teacher evaluation - Student achievement	+4					
HQ14d - Maths teacher evaluation - Peer review	-8		+4			
HQ16a - Difficulty in filling maths teacher vacancies						
HQ17a - Incentives to recruit or retain maths teachers					+6	

A positive relationship indicates that the variable is associated with higher attainment; a negative relationship that it is associated with lower attainment.

Shaded row(s) indicate a variable that was not statistically significant in any model.

Factors are presented first (questionnaire items included in each factor are indicated) followed by separate variables entered into the model.

Moderate or large quasi-effect sizes (5 points upwards) are more likely to indicate real effects; some smaller values might indicate borderline or spurious relationships, which can arise in any model.