A Cross-National Comparison of Intra-Class Correlation Coefficient in Educational Achievement Outcomes

Cengiz ZOPLUOGLU[*]

University of Minnesota

Abstract

The plausible range of intra-class correlation coefficient (ICC) is essential for both a priori sample size calculations in planning cluster-randomized trials and statistical adjustments of misaligned analysis of clustered data in meta-analytic studies. Recent efforts to create databases for ICC in educational achievement outcomes are based on the studies published only in the US and Europe. The current study aims to extend the existing information for the plausible range of ICC values in educational achievement outcomes to a global scale by examining the distributional characteristics of two-level unconditional ICC estimates across countries participating in two international studies, TIMSS and PIRLS. The findings suggest large variability in the unconditional ICC estimates across countries, and current standards do not apply to every country. Researchers should look for country-specific ICC estimates in planning cluster-randomized trials and in comparing studies across countries.

Key words: HLM, meta-analysis, intra-class correlation, multi-level modeling, TIMSS, PIRLS

Ozet

Kumeler-arasi (intra-class) korelasyon (KAK) katsayisinin olasi deger araliginin bilinmesi bireyler yerine kumelerin rastgele secilerek yapildigi arastirmalarda yeterli istatistiksel gucu elde etmek icin gerekli orneklem buyuklugunun arastirma oncesi hesaplamasinda kritik bir oneme sahiptir. Ayrica, meta-analitik calismalarda daha once yayinlanmis arastirmalarin verileri yorumlanirken, eger onceki arastirmalar hiyerarsik bir yapiya sahip veriyi dogrusal regresyona analizi ile analiz etmislerse, bu arastirmalarin sonuclarinin da istatistiksel olarak duzeltilmesi gerekmektedir. Bu duzeltme isleminde de, KAK katsayisinin tahmini deger araliginin bilinmesi gerekli olabilmektedir. Son zamanlarda, egitimde basariyi olcen test skorlari icin KAK katsayisinin tahmini deger araligini gosteren veri tabanlari olusturma cabalari sadece ABD ve baz i Avrupa ulkeleri ile sinirli kalmistir. Bu arastirma halihazirda egitimde basariyi olcen test skorlari icin KAK katsayisinin tahmini deger araligi hakkindaki bilgiyi kuresel olcekte genisletmeyi amaclamis, ve bu amacla iki uluslarasi calismanin, TIMSS ve PIRLS, verilerini kullanmistir. Arastirmani bulgulari KAK katsayisinin ulkeler arasinda onemli olcude degisebildigini gostermistir.

Anahtar Sozcukler: Hiyerarsik dogrusal modelleme, cok duzeyli modelleme, kumeler-arasi korelasyon, meta-analiz, TIMSS, PIRLS

[*] Ars. Gor., Minnesota Universitesi, Egitim Fakultesi, Egitim Psikolojisi Bolumu, Egitimde Nicel Yontemler Programi, zoplu001@umn.edu

Randomized trials have been increasingly used in the field of education since the 1990s to provide better evidence about how well new programs operate in classrooms, schools, and districts. The review of about 100 peer-reviewed journals between 1950 and 1997 based on the Campbell Collaboration database indicates that the number of randomized trials increased from a few tens to more than 2,000 in the field of education (Boruch, De Moya, & Snyder, 2002).

It is typical to assign individuals to the treatment groups in the randomized trials randomly, but this practice is not always viable for practical, ethical, logistic, administrative, and political reasons (Moerbeek, Van Breukelen, & Berger, 2008; Raudenbush, 1997). Moreover, it is sometimes preferable to randomly assign clusters to the treatments because of "having sizable spillover effects on individuals other than those who receive it" or "having the need to distinguish the treatment effect from cluster characteristics" (What Works Clearinghouse [WWC], 2005, p.4). Besides, social science data naturally have hierarchical structure, and the individuals are always nested in higher order clusters such as classrooms, schools, neighborhoods, and so on. So, it's a common practice to randomly assign the higher order clusters to the treatment groups, and to study the treatment effect in the higher order clusters (Raudenbush & Bryk, 2002).

Researchers are faced with two main issues in designing cluster-randomized trials due to the degree of similarity across individuals within a cluster measured by the intra-class correlation coefficient (ICC). The first issue is the reduced statistical power. The cluster-randomized trials require more subjects than the individual-randomized trials to obtain adequate statistical power, because the subjects within a cluster are not independent and do not bring unique information into the analysis. The second issue is the biased parameter estimates if the researchers decide to use traditional single level analysis for the clustered data. The clustered data violate the assumption of independence across observations within a cluster resulting to having biased standard error estimates. The first is a design issue and can be resolved prior to the research by computing the necessary sample size at each level to obtain the adequate statistical power and collecting enough data. The second is an analysis issue and can be resolved by using appropriate methods (e.g., multilevel models) for the clustered data. However, fitting multilevel models is not always a common practice for clustered data, and other researchers would like to correct the test statistics from misaligned analysis based on correction formulas, especially when combining the findings of other studies in meta-analytic research (Hedges, 2007).

ICC is one of the key elements required for both doing a priori sample size calculations in planning the cluster-randomized trials and adjusting the test statistics from the misaligned analysis of clustered data in meta analytic-studies (Turner, Prevost, &Thompson, 2004; Hedges, 2007; Hedges & Hedberg, 2007; Rotondi & Donner, 2009; Hedges & Rhoads, 2010). The plausible range of ICC values is crucial, and a priori guesstimate of ICC is mostly used in research design and meta-analytic studies. Recently, creating databases for the plausible ICC values in different measurement outcomes has been an appealing line of research to provide guidance for researchers and reviewers in different fields, because there is not always enough information in the literature to be able to guesstimate the ICC value. A review of multilevel reporting from 99 articles in 13 peer-reviewed journals indicates that only about 31% of the studies reported enough information to compute the ICC value (Dedrick, Ferron, Hess, Hogarty, Kromrey, Lang, Niles, & Lee, 2009). Similarly in other fields, there have been increasing efforts to create ICC databases for academic achievement outcomes to provide guidance for planning cluster-randomized experiments in education, and to synthesize the findings from previous educational research that used misaligned analysis of clustered data (Bosker & Witziers, 1996; Hedges & Hedberg, 2007; Stockford, 2009). The previous research reported an ICC value between 0.10 and 0.25 at the school level for educational achievement outcomes. What Works Clearinghouse (WWC) also uses the default value of .2 for achievement outcomes in their reviews to correct statistics from a misaligned analysis (WWC, 2008).

A limitation of the previous research that summarize the plausible range of ICC values for the educational achievement outcomes was including the studies published only in the US and Europe. It should be questioned whether the proposed standards for the ICC values in educational achievement outcomes reported by previous research are beneficial and relevant to scholars in other countries. The standards for the ICC values in education proposed by the previous research may mislead researchers in other countries in designing their experiments, because it is very likely that the variance attributed to the schools is different across countries with different educational, political, social, or economic characteristics.

Both a priori sample size calculations and adjustment formulas are sensitive even to small changes in ICC values, and small deviations may influence the researchers' decisions in the educational research design as well as in the comparison of studies across countries. The current standards for plausible ICC values in educational achievement outcomes should be used cautiously by researchers when planning multilevel studies in other countries or comparing studies across countries. It is necessary to extend the current knowledge about plausible ICC values for achievement outcomes to a global scale since there has not yet been a systematic effort to create an international database of the plausible ICC values for the educational achievement outcomes.

Using the data from two publicly available international datasets, TIMSS and PIRLS, the current study aims to extend the existing knowledge about plausible ICC values for educational achievement outcomes to a global scale. The study seeks to construct an international database for plausible ICC values in different achievement domains to provide guidance for the educational researchers in more than 80 countries. First, the unconditional ICC estimates at the school and country levels were derived from a three-level fully unconditional model. Second, the unconditional ICC for each country at different levels of grade, achievement domain, and study year was estimated from separate two-level unconditional models, and the distributional characteristics of the ICC estimates were examined. Third, the effects of possible influential variables on the ICC estimates were explored in a descriptive way.

Theoretical Framework

ICC and Statistical Power

Cluster-randomized trials have the advantages of "administrative efficiency, lessened risk of experimental contamination, and likely enhancement of subject compliance" (Donner & Clar, 2004, p.416). On the other hand, researchers pay an additional cost in the statistical power and precision because of the degree of similarity across individuals within a cluster. The subjects within a cluster are not generally independent from each other because of the shared experiences in the same environment or non-random assignment of the subjects into the clusters. The dependence within a cluster is measured by the ICC which is estimated by the proportion of variance between clusters to the total variance in the outcome. ICC values range from zero to one, and have substantial influence on the *design effect* for the mean which is formulated as

$$D = 1 + (n - 1) * \rho ,$$

where *n* is the number of individuals per cluster in a balanced design or average number of individuals per cluster in a non-balanced design, and *ρ* is the estimated ICC (Kish, 1965). The design effect is an important piece of information used in the research design for a priori sample size calculations in cluster-randomized trials. In the best scenario, an ICC of zero indicates that each individual in a cluster provides unique information into the analysis, and the design effect equals to one. So, the number of subjects required for the cluster randomization is equal to the number of subjects required for the individual randomization to obtain the adequate statistical power. On the other hand, the design effect is larger than one even for the small values of ICC. The design effect is equal to 1.95 for an ICC value of .05 and a cluster sample size of 20. This suggests that the researcher needs almost twice as many subjects from individual randomization to get the same statistical power in a cluster randomization for an ICC value of .05.

Reduced statistical power in cluster-randomized trials is one of the key design issues at the planning stage, and it can be resolved by computing the necessary sample sizes at each level prior to the research (Hedges & Rhoads, 2010; Moerbeek et al., 2008; Schochet, 2005). Four elements are required to conduct a priori sample size calculation in cluster-randomized trials: the desired level of significance for the statistical test, the expected effect size, the desired level of statistical power, and ICC. Theoretical formulas as well as the pre-prepared tables are available for cluster-randomized trials to carry out a priori sample size calculations based on these four elements (Hedges & Rhoads, 2010, Konstantopoulos, 2009). In addition, free softwares are available to researchers to carry out a priori sample size calculations in cluster-randomized trials (Spybrook, Raudenbush, Congdon, & Martinez, 2009; Rotondi, 2011). It is common to use the nominal alpha level of .05 and the power level of .80 in educational research for a priori sample size calculations, but the effect size and ICC should be estimated.

ICC and Type I Error Rate

Another consequence occurs when the clustered data are analyzed with a traditional single level analysis by ignoring the clustering. The single level analysis does not meet the assumption of independence in a cluster-randomized design, and ignoring this dependency may lead to biased parameter estimates, incorrect standard errors, and consequently incorrect statistical tests and effect sizes (Moerbeek et al., 2008). In a traditional fixed effect ANOVA analysis for a nested design, a nominal alpha level of .05 increased to the empirical values of .22 and .35 when the ICC was .05 and .10 respectively, and the average number of individuals is 30 in a cluster (Zucker, 1990). As a result of the simulation study, Kromrey and Dickinson (1996) also found that a nominal alpha level of .05 was inflated to .33 when the number of students per cluster was 30, and the ICC was .10. The most appropriate way to analyze clustered data is to use multilevel modeling, also known as Hierarchical Linear Modeling (Raudenbush & Byrk, 2002), that accounts for the variability among clusters in estimating the standard errors. Even though the use of multilevel modeling has been increasing, it is very common to see studies in the educational literature that ignore clustering in the statistical analysis. This fact is not just because of the researchers' ignorance but also because of the practical limitations. For instance, fitting multilevel models requires a sufficient number of clusters to model the variation among clusters accurately, and it is not very informative to fit a multilevel model with just five or six clusters. Or, an educational researcher who conducts an experiment with just two classes, one for control and another for experimental, cannot use multilevel modeling efficiently. However, the effect of clustering does not disappear in these cases, and the dependency between observations within a cluster still affects the test statistics, and inflates the Type I error rates. A formula for the adjustment of test statistics is proposed for studies that ignored the clustering in statistical analysis. Hedges (2007) derived a constant to compute the adjusted t statistic for clustering effect. The constant is computed as

$$c = \sqrt{\frac{(N-2) - 2\rho(n-1)}{(N-2)\,[1+\rho(n-1)]}},$$

where N is the number of total individuals in the study, n is the average number of individuals per cluster, ρ is the intra-class correlation. Then, the adjusted t statistic is calculated as

,

where t is the test statistic obtained from the single level analysis. The adjusted t statistic is evaluated based on the adjusted degrees of freedom computed as

$$h = \frac{[(N-2) - 2\rho(n-1)]^2}{(N-2)(1-\rho)^2 + n(N-2n)p^2 + 2(N-2n)\rho(1-\rho)} \; .$$

The correction formula requires a guesstimate of ICC to adjust the test statistics obtained from a misaligned analysis of the clustered data. These adjustments can also be used in meta-analytic studies when combining the result of other studies in the literature.

ICC and Influential Variables

The previous research suggested an average ICC of .19 for mathematics and .16 for language (Bosker & Witziers, 1996), an average of .22 for reading and mathematics (Hedges & Hedberg, 2007), and an average of .25 for mathematics, .19 for literacy, and 26 for science achievement scores (Stockford, 2009). Four main variables were examined as possible influential variables on ICC: grade level, achievement domain, socioeconomic status, and study year.

The results for the relationship between grade level and average two-level fully unconditional ICC values were mixed. In a meta-analytic study, Bosker and Witziers (1996) found a significant effect of grade level on the magnitude of the school effect. The average ICC for the secondary school level was .08 lower than the average ICC for the primary school level. Similarly, Hedges and Hedberg (2007) reported a negative correlation between grade level and average ICC. They found that the average ICC values decreased .005 per grade. However, Stockford (2009) found a non-linear relationship between grade level and average ICC. The average ICC for the primary school was slightly higher than the middle school level, and the average ICC for the middle school level was lower than the secondary school level.

Hedges and Hedberg (2007) reported a negative relationship between achievement level and the average two-level unconditional ICC estimates. The schools with low-achievement level had about 60% lower average ICC estimates for both mathematics and reading domains compared to all schools. Hedges and Hedberg (2007) also found that schools with low SES had about 11% lower average ICC in mathematics, and about 14% lower average ICC in science domains compared to all schools.

Purpose of the Study

A prior knowledge of plausible ICC values is essential for a priori sample size calculations in multilevel analysis and statistical adjustment of test statistics in single level analysis. A limitation of the previous research that created databases for the plausible ICC values in educational achievement outcomes was including the studies only published in the US or Europe, and those values may mislead researchers in other countries. The main purpose of the current study is to extend the existing information about plausible ICC values in the educational achievement outcomes to a global scale and to examine the distribution of the two-level fully unconditional ICC values across countries. Also, the effects of some previously studied influential variables on the average ICC values were examined.

Method

Sample and Instruments

The study used data from Trends in Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) developed by International Association for the Evaluation of Educational Achievement (IEA). TIMSS data were collected from both 4th and 8th grade students in 1995, 2003, 2007, and from only 8th grade students in 1999 on the mathematics and science achievement domains. The TIMSS datasets are publicly available for 84 countries and sub national education systems including the benchmark participants. PIRLS data were collected from only 4th grade students in 2001 and 2006 on the reading achievement domain. The PIRLS datasets are also publicly available for 52 countries and sub national education systems including the benchmark participants. The countries that participated in these studies were summarized at Table 1 and Table 2. The TIMSS and PIRLS used two-stage stratified cluster sampling designs. First, schools were sampled with probability proportional to size. Then, one or more intact classes of students were sampled from the target grades. In general, there were one or two classrooms sampled from each school. The target populations were 4th and 8th grade students defined by UNESCO's International Standard Classification of Education (ISCED). Fourth grade population included the students enrolled in the grade that represented four years of formal schooling, and 8th grade population included the students enrolled in the grade that represented eight years of formal schooling.

The numbers of items administered in TIMSS and PIRLS were also reported in Table 1 and Table 2. The students took a small subset of items in different booklets and a complex scaling procedure was carried out to produce the achievement scores. The process included calibrating the achievement test items, creating principal components from the student questionnaire data for use in conditioning, generating IRT scale scores for overall mathematics, science, and reading proficiency, and placing the proficiency scores on the metric used to report the results from previous years. All scores were placed on the metric used in 1995 to be able to make comparisons across years. Item Response Theory (Lord, 1980) and plausible value technology (Rubin, 1987) were used to calibrate students' ability level estimates. Plausible value technology allowed more accurate estimation of student ability by taking a student's background information into account, and reduced the measurement error because only a subset of items were administered to a student. In TIMSS and PIRLS, five separate estimates of achievement scores that represented the likely distribution of a student's proficiency in a domain were provided. The plausible values were put on a scale with a mean of 500 and standard deviation of 100.

Both studies have complicated sampling, instrumentation, administration, and scaling procedures that are beyond the purpose of the current study. The readers are encouraged to review the technical manuals of these studies to learn more details about sampling, instrumentation, administration, and scoring procedures (Martin & Kelly, 1996; Martin, Gregory, & Stemler, 2000; Martin, Mullis, & Kennedy, 2003; Martin, Mullis, & Chrostowski, 2004; Martin, Mullis, & Kennedy, 2007; Olson, Martin, & Mullis, 2008).

Data Analysis

The TIMSS and PIRLS datasets have a three-level structure. The students are nested within classrooms/schools, and the classrooms/schools are nested within countries. It is also possible to think of the design structure as a four-level structure with nested classrooms in schools. However, there are just one classroom sampled in most schools in TIMSS and PIRLS, and it is not possible to model the variability among classrooms within schools. So, a three level structure was used in the current research.

A random effects ANOVA model, also known as the fully unconditional model, is the simplest multilevel model and does not include any covariate at any level. The fully unconditional model is generally used as the first step in multilevel analysis. The magnitude of ICC estimated from the fully unconditional model is used to decide whether or not the multilevel modeling is required for the statistical analysis. In a three-level fully unconditional model, the variance in the achievement outcome can be partitioned into three components: among students within schools at level 1, among schools within countries at level 2, and among countries at level 3 (Raudenbush & Bryk, 2002). In the current study, a three level fully unconditional model was first fitted to each TIMSS and PIRLS dataset currently available. The proportion of variability accounted by schools and countries were estimated. The fully unconditional three-level model is written for the *i*th student in the *j*th school in the *k*th country at Level 1 as

$$\mathbf{Y}_{\mathbf{ijk}} = \mathbf{0jk} \qquad \mathbf{ijk}$$

,

for the *j*th school in the *k*th country at Level2 as

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

and for the *k*th country at Level 3 as

$\beta_{00k} = \gamma_{000} + \mu_{00k}$,

where e_{ijk} is a random student effect (level 1 residual), r_{0jk} is a random school effect (level 2 residual), and μ_{00k} is a random country effect (level 3 residual). The variance components among students within schools, among schools within countries, and among countries are symbolized by $\sigma^2, \tau_\pi, \tau_\beta$ respectively. Finally, the unconditional ICC at Level 2 and Level 3 are calculated as the following:

$$\text{ICC (Level 2} = \tau\downarrow\pi/(\sigma\uparrow 2 + \tau\downarrow\pi + \tau\downarrow\beta)\ \text{,the proportion of variance among schools within countr}$$

$$\text{ICC (Level 3)} = \frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta}\ \text{,the proportion of variance among countries}$$

Another purpose of the current study is to examine the distribution of ICCs estimated from two-level unconditional models across countries. So, a two-level fully unconditional model was fitted to each country's data separately within every TIMSS and PIRLS dataset available. In a two-level fully unconditional model, the variance in achievement outcomes can be partitioned into two components: among students within schools at level 1, and among schools at level 2. In a similar way, the fully unconditional two-level model is written for the *i*th student in the *j*th school at Level 1 as

$Y_{ij} =$ [illegible] $+$ [illegible]

,

for the *j*th school at Level2 model as

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

where e_{ij} is a random student effect (level 1 residual), and r_{0j} is a random school effect (level 2 residual). The variance components among students within schools and among schools are symbolized by σ^2, τ_π respectively. Finally, the unconditional ICC at Level 2 for a country can be estimated as the following:

$$\text{ICC} = \frac{\tau_\pi}{\sigma^2 + \tau_\pi}\text{, the proportion of variance among schools}$$

HLM 6.2 (Raudenbush, Bryk , & Congdon, 2005) was used to estimate the variance components in the three-level and two-level fully unconditional models. The dependent variables were plausible values for mathematics, science, and reading achievement available in TIMSS 1995, 1999, 2003, 2007, and in PIRLS 2001, 2006 datasets. There were five plausible values assigned to every student for each domain in TIMSS and PIRLS datasets. HLM 6.2 estimates the variance components for each plausible value separately and then reports the average for each variance component. The weights that are available in the datasets were also used in the statistical analysis. The total student weight (TOTWGT) and school weight (SCHWGT) were used at level 1 and level 2 in the analysis. There was no weight available at level 3 in the datasets.

The variance components at Level 2 and Level 3 obtained from the three-level fully unconditional model and the distributional characteristics of ICCs estimated from the two-level fully unconditional model across countries were reported at different levels of grade, achievement domain, and study year.

Results

The variance components estimated from the three-level fully unconditional model are reported in Table 3 and Table 4, and visualized in Figure 1. In general, the total variability in the achievement outcomes due to the school and country characteristics was above 60% except in 1995. The total variability accounted by both school and country characteristics increased from about 45% to about 75% for fourth graders and from about 40% to about 70% for eighth graders across years. The total variability accounted by both schools and countries showed a similar pattern across different achievement domains.

The variability accounted by only school characteristics (ICC at Level2) did not increase substantially from 1995 to 2007. It was about 20% for the fourth graders and about 25% for the eighth graders, and a similar pattern was observed for different achievement domains. However, the variability accounted by only country characteristics (ICC at Level 3) substantially increased from about 30% to about 50% for the fourth graders and from about 20% to about 40% for the eighth graders across years.

Table 5 presents the summary statistics for the distributions of the two-level fully unconditional ICC estimates across countries at different levels of grade, achievement domain, and study year. The average ICC estimate at the 8th grade level was about .33 for mathematics, and about .29 for science achievement domains. The average ICC estimate at the 4th grade level was .25 for mathematics, .24 for science, and .25 for reading achievement domains. The current study found slightly higher average two-level fully unconditional ICC estimates compared to the standards in the literature. The previous research reported an average ICC between .10 and .25 for educational achievement outcomes (Bosker & Witziers, 1996; Hedges & Hedberg, 2007; Stockford, 2009).

There was substantial variability across countries in the two-level fully unconditional ICCs. While the proportion of variance in achievement outcomes due to schools was lower than .10 for some countries, it was higher than .60 for some other countries. In most occasions, the two-level unconditional ICCs were distributed with a standard deviation between .12 and .16 regardless of the achievement domain, year, and grade level. The only exception was the 4th grade data in 1995. The ICCs were distributed with a standard deviation of .07 for mathematics achievement outcome and .08 for science achievement outcome around the mean. The skewness values were not bigger than .75 and the kurtosis values were between 2.5 and 2.8 for most occasions. The histograms for the distributions of the two-level fully unconditional ICCs at different levels of achievement domain, grade level, and study year are reported in Appendix A. The exact two-level fully unconditional ICC estimates for each country are reported in Appendix B at the different levels of achievement domains, study year, and grade level.

The change in average two-level unconditional ICC estimates across years was shown in Figure 2. The average two-level fully unconditional ICC estimates at the 8th grade level did not change much across years for both mathematics and science achievement domains. However, it slightly increased at the 4th grade level for all types of achievement domains. The figure also indicated that the average two-level unconditional ICC estimate for mathematics was somewhat higher than the other achievement domains. The difference was more apparent at the 8th grade level and ignorable at the 4th grade level. This finding was consistent with the previous research.

The current study found that the average two-level fully unconditional ICC estimate was slightly higher at the 8th grade level than at the 4th grade level. The average two-level fully unconditional ICC estimate at the 8th grade level was 30% higher for mathematics, and 20% higher for science domains than the average two-level fully unconditional ICC estimate at the 4th grade level, but the difference got smaller across years. The finding about the relationship between grade level and ICC was inconsistent with the previous research. While the previous research suggested a decrease in ICC for an increase in grade level, the current study found a slightly positive relationship.

A detailed examination of Appendix B suggested that the ICC estimates were very similar across different years and achievement domains within a country. So, the ICC estimates were averaged across the years and achievement domains within a country and the countries were classified based on their average ICC estimates in Table 6 and 7 for different grade levels.

In Table 8 and 9, the countries were put into four different achievement levels, and the average ICC estimate across countries within each achievement level was computed to examine the relationship between the achievement level and the two-level fully unconditional ICC estimates. The results did not suggest a clear-cut evidence in most situations to argue that there is a systematic relationship between the average achievement score and average two-level fully unconditional ICC estimate. In some cases, the average ICC estimate was lower for high achieving countries. For instance, the average ICC estimate decreased from about .40 to about .24 for science and mathematics domain in 2003 and 2007, and from about .3 to .2 for reading domain in 2006 at 4th grade level as the achievement level increased.

The gross domestic product produced per capita (GDP) in the units of US dollars was obtained for each country as a measure of socioeconomic status (SES) to examine the relationship between country SES and two-level fully unconditional ICC estimates. The GDP data were available between 1999 and 2007 for 79 countries on which the unconditional ICCs were estimated (UNESCO, 2010). Similar to the achievement, the countries were put into four different GDP levels, and the average two-level fully unconditional ICC estimate across the countries within each GDP level was computed. The results reported in Table 10 and 11 did not suggest a systematic relationship between country SES and two-level fully unconditional ICCs at the 8th grade level. However, the average two-level fully unconditional ICC estimate had a tendency to be lower for the countries with higher GDP for mathematics and science achievement domain in 2003 and 2007, and for reading achievement domain in 2001 and 2006 at the 4th grade level.

Discussion

The key finding was that the variance in the educational achievement outcomes attributed to schools has an important variability across countries. The two-level fully unconditional ICC estimates were between .05 and .70 across the countries. This finding suggests that researchers in some countries may have underestimated the necessary sample size while researchers in some other countries may have overestimated the necessary sample size to obtain adequate statistical power at the planning stage if the standards suggested in the literature for plausible range of ICC values based on the US and Europe were used. The researchers in other countries should look for country-specific ICC estimates other than using the standard values proposed in the literature when planning cluster-randomized trials in educational research.

The previous research proposed a plausible range between .10 and .25 for the two-level unconditional ICC estimates for the achievement outcomes in the US. But, the TIMSS and PIRLS datasets suggest slightly higher values. Examination of the US data revealed that the average two-level unconditional ICC estimate across years was .37 for both mathematics and science achievement domains at the 8th grade level, and .29 for mathematics, .33 for science, and .25 for reading achievement domains at the 4th grade level. This disparity may result in different applications in practice when designing cluster-randomized trials in education. For instance, the design effect for an average cluster size of 30 is 6.8, 9.7, and 10.15 with an ICC of .2, .3 and .35 respectively, indicating that the necessary sample size should be computed 42% or 49% higher for the US if the plausible values suggested by TIMSS and PIRLS were used.

Variability in the two-level fully unconditional ICC estimates across countries also implies that researchers should be cautious when doing cross-national comparisons. For instance, combining the effect sizes from studies conducted in different countries in meta-analytic studies may not be suitable. A country-specific ICC value is necessary to adjust the test statistics, and to re-compute the effect sizes for the studies conducted in different countries before any comparison. The values reported in this study are intended to provide an initial estimate of ICC in different countries to be used for cross-national comparisons and analysis.

A detailed examination of exact ICC estimates in Appendix B show that the two-level unconditional ICC estimates were consistently higher for some countries while they are consistently lower for some other countries. For instance, the ICC estimates were below .15 for Korea, Cyprus, Norway, and Slovenia, and above .4 for Dubai, Colombia, Lebanon, and Malaysia in all occasions. This finding may recommend an important line of research for future investigators to explore why some countries have lower school effects while other countries have higher school effects. Which country level characteristics have influence on the magnitude of unconditional ICC estimates? The findings in the current research do not support clear-cut evidence that the magnitude of school effect has a strong systematic relationship with the country-level achievement or country-level SES. Future research may address some country characteristics to explain the variability in school effects across countries.

References

Boruch, R., De Moya, D., & Snyder, B. (2002). The Importance of Randomized Field Trials in Education and Related Areas. In F. Mosteller, & R. Boruch, *Evidence Matters: Randomized Trials in Education Research* (pp. 59-70). Washington, DC: The Brooking Institution.

Boruch, R., May, H., Turner, H., Lavenberg, J., Petrosino, A., de Moya, D., et al. (2004). Estimating the effects of interventions that are deployed in many places : Place- Randomized Trials. *American Behavioral Scientist , 47*, 608-633.

Bosker, R. J., & Witziers, B. (1996). *A Meta Analytical Approach Regarding School Effectiveness: The True Size of School Effects and the Effect Size of Educational Leadership (ERIC Document Reproduction Service No. ED392147).*

Chudgar, A., & Luschei, T. F. (2009). National Income, Income Equality, and the Importance of Schools: A Hierarchical Cross-National Comparison. *American Educational Research Journal, 46(3)*, 626-658.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., et al. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *REVIEW OF EDUCATIONAL RESEARCH , 79*, 69-102.

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health , 94* (3), 416-422.

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics , 32*, 151-179.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group- Randomized Trials in Education. *Educational Evaluation And Policy Analysis , 29*, 60-87.

Hedges, L., & Rhoads, C. (2009). *Statistical Power Analysis in Education Research (NCSER 2010-3006).* Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Klar, N., & Donner, A. (2001). Current and future challenges in the design and analysis of cluster randomized trials. *Statistics in Medicine , 20*, 3729-3740.

Konstantopoulos, S. (2009). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. *Evaluation Review , 33* (4), 335-357.

Kromrey, J. D., & Dickinson, W. B. (1996). Detecting unit of analysis problems in nested designs: Statistical power and type I error rates of the F test for groups-within treatments effects. *Educational and Psychological Measurement , 56*, 215-231.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Martin, M. O., Gregory, K. D., & Stemler, S. E. (2000). *TIMSS 1999 Technical Report.* Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V., & Kennedy, A. M. (2003). *PIRLS 2001 Technical Report.* Chestnut Hill, MA: Boston College.

Martin, M., & Kelly, D. (1996). *TIMSS 1995 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M., Mullis, I., & Chrostowski, S. (2004). *TIMSS 2003 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston Center.

Martin, M., Mullis, I., & Kennedy, A. (2007). *PIRLS 2006 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Moerbeek, M., Van Breukelen, G. J., & Berger, M. B. (2008). Optimal Designs for Multilevel Studies. In J. d. Leeuw, & M. Erik, *Handbook of Multilevel Analysis* (pp. 177-206). New York, NY: Springer.

Murray, D. M., & Blitstein, J. L. (2003). Methods of reduce the impact of intraclass correlationin group-randomized trials. *Evaluation Review , 27*, 79-103.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review , 27* (1), 79-103.

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

Pong, S.-l., & Pallas, A. (2001). Class Size and Eighth-Grade Math Achievement in the United States and Abroad. *Educational Evaluation and Policy Analysis , 23* (3), 251-273.

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods , 2* (2), 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Application and Data Analysis Methods. 2nd Edition.* Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2005). *HLM 6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International.

Rotondi, M. A. (2011). CRTSize: Sample Size Estimation Functions for Cluster Randomized Trials. R package version 0.2. http://CRAN.R-project.org/package=CRTSize.

Rotondi, M. A., & Donner, A. (2009). Sample Size Estimation in Cluster Randomized Educational Trials: An Empirical Bayes Approach. *Journal of Educational and Behavioral Statistics , 34* (2), 229-237.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Schochet, P. Z. (2005). *Statistical power for random assignment: Evaluations of education programs.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software V.2.0. Available at www.wtgrantfoundation.org.

Stockford, S. M. (2009). Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement (Doctoral dissertation, Arizona State University, 1990). Dissertation Abstracts International,70,08.

Turner, R. M., Prevost, A. T., & Thompson, S. G. (2004). Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine , 23*, 1195-1214.

UNESCO. (2010). *The Gross Domestic Product Per Capita Custom Table.* Retrieved March 26, 2011, from UNESCO Institute for Statistics Data Center: http://stats.uis.unesco.org/unesco/TableViewer/document.aspx?ReportId=143&IF_Language=eng

WhatWorksClearinghouse. *Key Items To Get Right When Conducting a Randomized Controlled Trial in Education.* Washington, DC: U.S. Department of Educations, Institute of Education Sciences. Retrieved March 26, 2011, http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf.

WhatWorksClearinghouse. (2008). *Procedures and Standards Handbook Version 2.* Washington, DC: U.S. Department of Educations, Institute of Education Sciences. Retrieved March 26, 2011, http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf.

Zucker, D. M. (1990). An analysis of variance pitfall: The fixed effects analysis in a nested design. *Educational and Psychological Measurement , 50*, 731-738.

Table 1. The Summary Information of TIMSS Datasets

Year		4th	8th	Year		4th	8th
1995	Number of Countries	25	40	2003	Number of Countries	29	51
	Average Number of Schools (per country)	161.0	156.1		Average Number of Schools (per country)	160.1	151.9
	Average Total Number of Students (per country)	6798.5	6799.8		Average Total Number of Students (per country)	4410.2	4663.3
	Number of Math. Items	102	151		Number of Math. Items	161	194
	Number of Sci. Items	97	135		Number of Sci. Items	152	189
1999	Number of Countries	-	38	2007	Number of Countries	44	57
	Average Number of Schools (per country)	-	156.9		Average Number of Schools (per country)	156.5	144.4
	Average Total Number of Students (per country)	-	4631.2		Average Total Number of Students (per country)	4253.7	4284.5

Number of Math. Items	-	162	Number of Math. Items	192	238
Number of Sci. Items	-	146	Number of Sci. Items	194	240

Table 2. The Summary Information of PIRLS Datasets

	2001	2006
Number of Countries	35	45
Average Number of Schools (per country)	154.7	169.5
Average Total Number of Students (per country)	4069.2	4780.8
Number of Reading Items	98	126

Table 3. The Variance Components Estimated from Three-Level Fully Unconditional Model for Science (TIMSS) and Reading (PIRLS) Domains

Domain	SCIENCE							READING	
Year	2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	4th	4th
k	245,172	187,673	237,833	127,896	180,700	271,964	176,709	215,137	146,490
j	8,265	6,902	7,751	4,642	6,071	6,245	4,206	5,570	5,570
i	57	44	51	29	38	40	26	36	36
σ^2	4098.5	4552.5	4434.6	5918.4	5422.3	5604.3	6332.8	3601.3	3577.7
τ_π	2568.7	3742.5	2822.8	3274.7	3200.2	2550.0	1881.5[a]	2609.3	2664.1
τ_β	3360.8	8378.1	6485.6	9414.1	4692.5	1309.0	3641.3	5471.8	4146.9
ICC (Level 2)	.26	.22	.21	.18	.24	.27	.16	.22	.26
ICC (Level 3)	.34	.50	.47	.51	.35	.14	.31	.47	.40
Total ICC	.60	.72	.68	.69	.59	.41	.47	.69	.66

Notes. k: number of students, j: number of schools, i: number of countries, σ^2: the variability among students within schools, τ_π : the variability among schools within countries, τ_β : the variability among countries, ICC: intra-class correlation.

[a] The variance components are not significant at α=.05

Table 4. The Variance Components Estimated from Three-Level Fully Unconditional Model for Mathematics (TIMSS) Domain

Year	2007		2003		1999	1995	
Grade	8th	4th	8th	4th	8th	8th	4th
k	245,172	187,673	237,833	127,896	180,700	271,964	176,709

j	8,265	6,902	7,751	4,642	6,071	6,245	4,206
i	57	44	51	29	38	40	26
σ^2	4294.7	3849.9	4072.0	4425.6	4893.1	4747.5	5862.4
τ_π	2988.8	3224.8	2888.2	2554.8	2941.5	2349.4	1576.2[a]
τ_β	4626.9	8384.6	6173.5	7537.8	4993.0	2407.6	2593.5
ICC (Level 2)	.25	.21	.22	.18	.23	.25	.16
ICC (Level 3)	.39	.54	.47	.52	.39	.25	.26
Total ICC	.64	.75	.69	.70	.62	.50	.42

Notes. k: number of students, j: number of schools, i: number of countries, σ^2: the variability among students within schools, τ_π : the variability among schools within countries, τ_β : the variability among countries, ICC: intra-class correlation.

[a] The variance components are not significant at α=.05

Table 5. The Summary Statistics for the Two-Level Fully Unconditional Intra-Class Correlation Coefficients across Countries

Domain	Mathematics							Science							Reading	
Year	2007		2003		1999	1995		2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	8th	4th	8th	4th	8th	8th	4th	4th	4th
N	57	44	51	29	38	40	26	57	44	51	29	38	40	26	36	36
Mean	.31	.27	.33	.27	.35	.30	.19	.29	.27	.29	.24	.31	.27	.19	.23	.28
SD	.14	.14	.16	.12	.16	.14	.07	.13	.13	.14	.12	.15	.12	.08	.11	.14

Min	.03	.07	.06	.04	.07	.10	.05	.06	.08	.08	.04	.06	.10	.04	.08	.08
Max	.65	.62	.71	.56	.71	.63	.34	.61	.58	.61	.51	.75	.54	.35	.52	.65
Skewness	.38	.74	.54	.35	.17	.64	-.14	.37	.51	.38	.31	.44	.50	.03	.78	.65
Kurtosis	2.80	2.81	2.49	2.56	2.58	2.63	2.28	2.70	2.55	2.27	2.18	3.2	2.33	2.26	2.55	2.70

Notes. N represents the number of countries for which separate two-level fully unconditional models were run.

Table 6. The Classification of Countries based on Their Two-Level Fully Unconditional ICC values at 8th Grade

ICC Level	Countries
< 0.1	Cyprus, Algeria, Korea, Slovenia
0.1 – 0.2	Bahrain, Bosnia, Denmark, Estonia, Finland, Iceland, Japan, Latvia, Morocco, Norway, Portugal, Palestinian National Authority, Serbia, Tunisia
0.2 – 0.3	Armenia, Spain(Basque Country), Botswana, Canada, Canada (Ontario), Czech Republic, Egypt, Spain, France, Georgia, Greece, Hungary, Iran, Italy, Jordan, Kuwait, Lithuania, Mongolia, Oman, Saudi Arabia, Slovak Republic, Sweden, Syria, Chinese Taipei, United States(Indiana), Ukraine, United States (Minnesota)
0.3 – 0.4	Austria, Belgium(French), Bulgaria, Chile, Canada (Quebec), Ghana, Ireland, Israel, Moldova, Macedonia, Qatar, Romania, Russian Federation, Scotland, El Salvador, Thailand, Turkiye, United States (Massachusetts), United States
0.4 – 0.5	Dubai, Australia, Belgium(Flemish), Canada(British Columbia), Colombia, England, Indonesia, Lebanon, New Zealand
> 0.5	Switzerland, Germany, Hong Kong, Malta, Malaysia, Netherlands, Philippines, Singapore, South Africa

Table 7. The Classification of Countries based on Their Two-Level Fully Unconditional ICC values at 4th Grade

ICC Level	Countries
< 0.1	Iceland, Japan, Korea, Norway, Poland, Slovenia
0.1 – 0.2	Austria, Belgium(Flemish), Belgium(French), Bosnia, Canada, Canada(Novia Scotia), Canada (Quebec), Czech Republic, Cyprus, Denmark, France, Ireland, Lithuania, Luxembourg, Netherlands, Palestinian National Authority, Serbia, Scotland, Sweden, Chinese Taipei, United States(Indiana), Ukraine
0.2 – 0.3	Australia, Bahrain, Spain(Basque Country), Botswana, Canada(Alberta), Canada(British Columbia), Canada (Ontario), Algeria, Egypt, England, Spain, Greece, Hong Kong, Hungary, Israel, Italy, Jordan, Kuwait, Latvia, Mongolia, New Zealand, Oman, Portugal, Qatar, Saudi Arabia, Syria, Tunisia, United States (Massachusetts), United States (Minnesota), United States
0.3 – 0.4	Armenia, Bulgaria, Germany, Georgia, Iran, Moldova, Romania, Singapore, El Salvador, Slovak Republic, Thailand, Trinidad And Tobago, Turkiye
0.4 – 0.5	Dubai, Argentina, Belize, Colombia, Ghana, Indonesia, Kazakhstan, Macedonia, Philippines, Russian Federation, Yemen
> 0.5	Lebanon, Morocco, Malta, Malaysia

Table 8. Mean Two-Level Fully Unconditional ICCs by Achievement Level for Science and Reading Domains

Domain		Science							Reading	
Year		2007		2003		1999	1995		2006	2001
Grade		8th	4th	8th	4th	8th	8th	4th	4th	4th
Level 1 (Score < 300)	N	-	3	2	1	1	-	-	1	-
	Mean Score	-	266.437 (27.58)	220.614 (12.60)	243.443	224.95	-	-	287.61	-
	Mean ICC	-	.435 (.090)	.442 (.127)	.390	.568	-	-	.614	-
Level 2 (300<Score<400)	N	4	5	5	3	2	1	-	4	4
	Mean Score	343.95 (14.21)	355.86 (15.20)	385.39 (5.27)	305.515 (1.43)	334.90 (13.75)	370.25	-	350.154 (17.98)	370.56 (17.97)
	Mean ICC	.358 (.046)	.354 (.032)	.321 (.071)	.358 (.013)	.293 (.234)	.468	-	.324 (.085)	.447 (.072)
Level 3 (400<Score<500)	N	34	7	20	6	16	26	11	8	8
	Mean Score	448.29 (5.30)	449.01 (12.31)	452.56 (6.89)	459.40 (14.46)	455.12 (6.36)	467.06 (4.99)	470.27 (8.54)	464.43 (12.86)	460.10 (12.22)
	Mean ICC	.286 (.023)	.264 (.046)	.237 (.025)	.208 (.047)	.298 (.026)	.269 (.023)	.195 (.028)	.307 (.042)	.296 (.061)
Level 4 (Score<500)	N	19	29	24	19	19	13	14	32	24
	Mean Score	529.05 (3.96)	531.49 (3.17)	530.11 (3.97)	527.446 (3.97)	534.19 (3.61)	520.20 (3.52)	527.58 (4.68)	536.41 (2.64)	530.93 (2.90)
	Mean ICC	.296 (.033)	.237 (.022)	.314 (.031)	.227 (.027)	.304 (.040)	.269 (.035)	.185 (.021)	.207 (.018)	.249 (.024)

Notes. The numbers in parantheses are standard errors

Table 9. Mean Two-Level Fully Unconditional ICCs by Achievement Level for Mathematics Domain

Year		2007		2003		1999	1995	
Grade		8th	4th	8th	4th	8th	8th	4th
Level 1 (Score < 300)	N	-	2	2	1	1	-	-
	Mean Score	-	266.14 (28.79)	247.14 (10.99)	272.53	260.057	-	-
	Mean ICC	-	.317 (.129)	.469 (.136)	.422	.485	-	-
Level 2 (300<Score<400)	N	15	7	8	4	4	3	-
	Mean Score	361.65 (7.58)	344.108 (11.27)	372.28 (7.86)	351.06 (10.76)	360.52 (11.43)	363.29 (12.95)	-
	Mean ICC	.295 (.029)	.396 (.048)	.329 (.067)	.381 (.034)	.339 (.093)	.364 (.137)	-
Level 3 (400<Score<500)	N	26	10	24	7	15	23	11
	Mean Score	456.086 (5.51)	471.01 (7.78)	461.85 (6.66)	478.23 (7.60)	457.48 (6.72)	472.92 (4.42)	463.74 (7.49)
	Mean ICC	.314 (.031)	.246 (.042)	.282 (.025)	.234 (.042)	.329 (.030)	.277 (.024)	.207 (.024)
Level 4 (Score<500)	N	16	25	17	17	18	14	14
	Mean Score	531.67 (8.11)	531.346 (5.66)	534.52 (7.77)	533.12 (5.89)	537.55 (7.31)	527.77 (7.21)	533.09 (6.51)
	Mean ICC	.324 (.037)	.242 (.025)	.377 (.043)	.244 (.030)	.365 (.047)	.327 (.045)	.171 (.017)

Notes. The numbers in parantheses are standard errors

Table 10. Mean Two-Level Fully Unconditional ICCs by Gross Domestic Product (GDP) Levels for Science and Reading Domains

Domain	Science					Reading	
Year	2007		2003		1999	2006	2001
Grade	8th	4th	8th	4th	8th	4th	4th

Level 1 (GDP < 10,000)	N	17	11	18	9	19	6	13
	Mean GDP	5,636$ (477.98)	5,706$ (597.10)	4,939$ (614.79)	4,275$ (908.04)	5,147$ (543.02)	4,535$ (853.85)	6,457$ (647.29)
	Mean ICC	.278 (.026)	.343 (.043)	.326 (.029)	.327 (.030)	.324 (.030)	.352 (.058)	.398 (.033)
Level 2 (10,000<GDP<25,000)	N	13	8	15	6	13	11	8
	Mean GDP	16,144$ (1337.10)	17,312$ (1636.71)	16,413$ (1347.92)	17,259$ (2161.64)	18,977$ (1420.86)	16,172$ (1466.49)	18,285$ (1434.06)
	Mean ICC	.310 (.036)	.317 (.053)	.226 (.032)	.197 (.035)	.234 (.037)	.301 (.035)	.221 (.034)
Level 3 (25,000<GDP<40,000)	N	10	10	11	10	6	14	11
	Mean GDP	31,683$ (1406.86)	34,134$ (1260.25)	31,454$ (1158.87)	31,593$ (1271.95)	27,774$ (1186.34)	33,533$ (1080.08)	30,084$ (1035.81)
	Mean ICC	.224 (.049)	.194 (.031)	.362 (.052)	.226 (.044)	.414 (.083)	.179 (.020)	.227 (.042)
Level 4 (GDP > 40,000)	N	6	7	-	-	-	6	-
	Mean GDP	50,202$ (2183.32)	48,819$ (2306.01)	-	-	-	56,277$ (5448.99)	-
	Mean ICC	.360 (.075)	.260 (.035)	-	-	-	.183 (.020)	-

Notes. The numbers in parantheses are standard error

Table 11. Mean Two-Level Fully Unconditional ICCs by Gross Domestic Product (GDP) Levels for Mathematics Domain

Year		2007		2003		1999
Grade		8th	4th	8th	4th	8th
Level 1 (GDP < 10,000)	N	17	11	18	9	19
	Mean GDP	5,636$ (477.98)	5,706$ (597.10)	4,939$ (614.79)	4,275$ (908.04)	5,147$ (543.02)
	Mean ICC	.303 (.026)	.359 (.042)	.347 (.032)	.373 (.024)	.353 (.030)
Level 2 (10,000<GDP<25,000)	N	13	8	15	6	13
	Mean GDP	16,144$ (1337.10)	17,312$ (1636.71)	16,413$ (1347.92)	17,259$ (2161.64)	18,977$ (1420.86)
	Mean ICC	.325 (.032)	.335 (.056)	.269 (.036)	.211 (.032)	.283 (.045)
Level 3 (25,000<GDP<40,000)	N	10	10	11	10	6
	Mean GDP	31,683$ (1406.86)	34,134$ (1260.25)	31,454$ (1158.87)	31,593$ (1271.95)	27,774$ (1186.34)
	Mean ICC	.240 (.057)	.196 (.035)	.418 (.059)	.245 (.046)	.491 (.078)
Level 4 (GDP > 40,000)	N	6	7	-	-	-
	Mean GDP	50,202$ (2183.32)	48,819$ (2306.01)	-	-	-
	Mean ICC	.379 (.087)	.264 (.043)	-	-	-

Notes. The numbers in parantheses are standard error

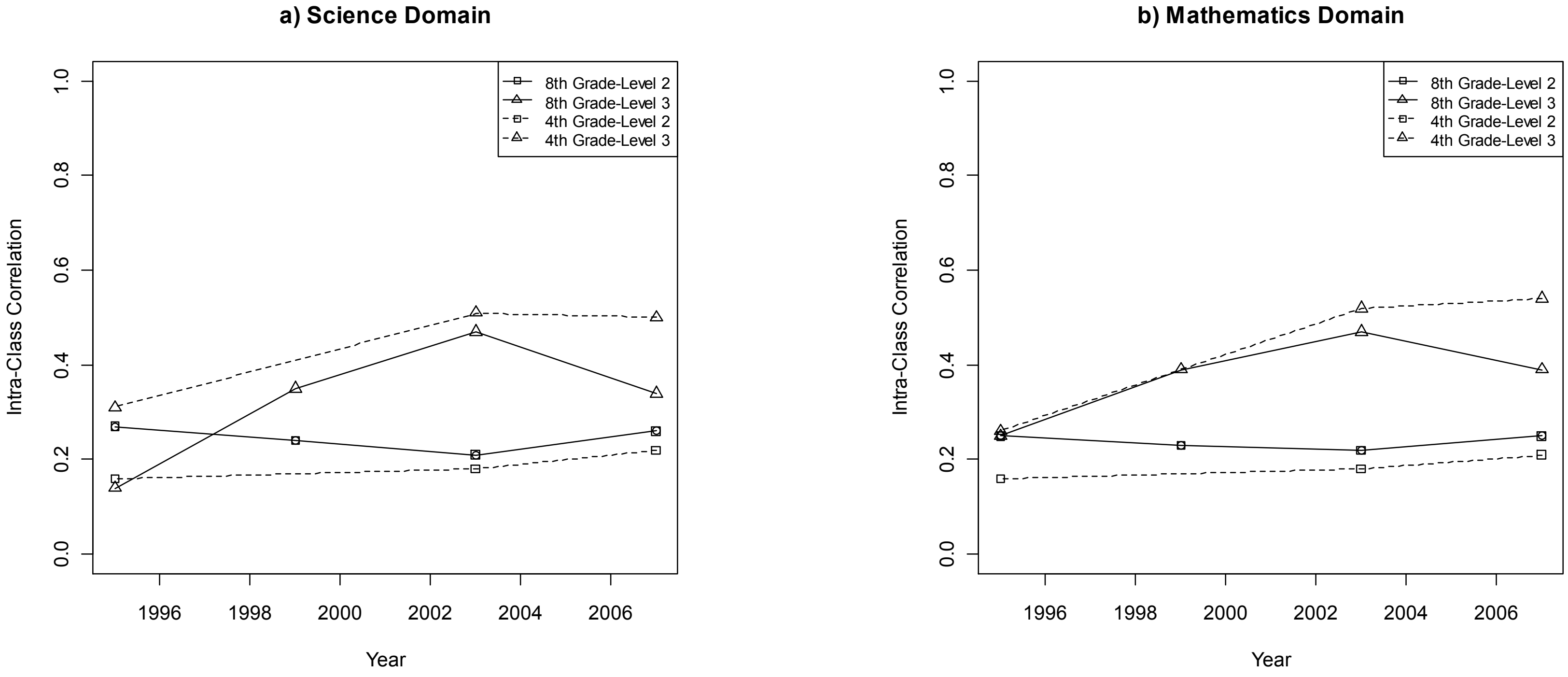


Figure 1. The Intra-Class Correlation Coefficients Estimated From Three-Level Fully Unconditional Model for Science and Mathematics Achievement Domain

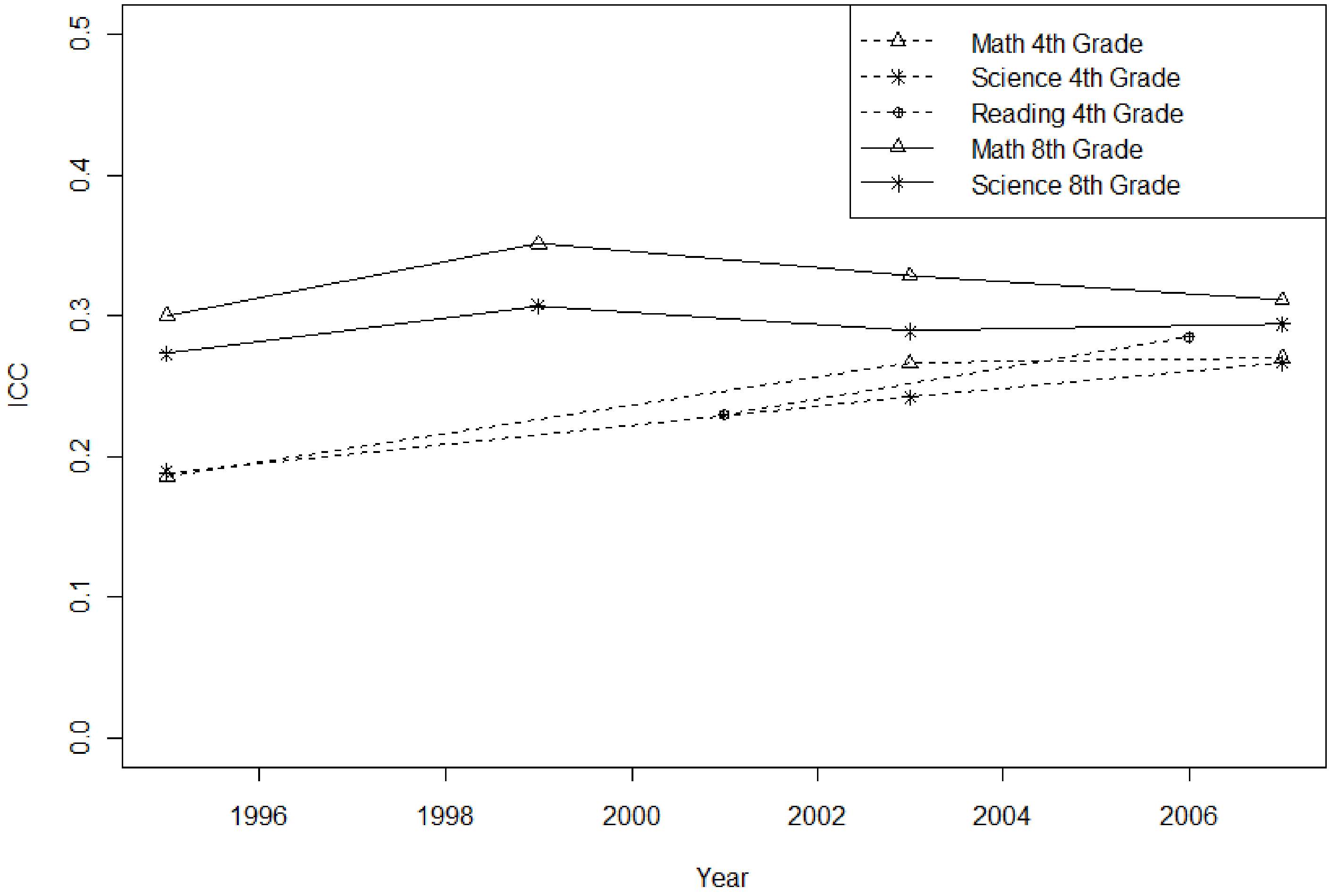
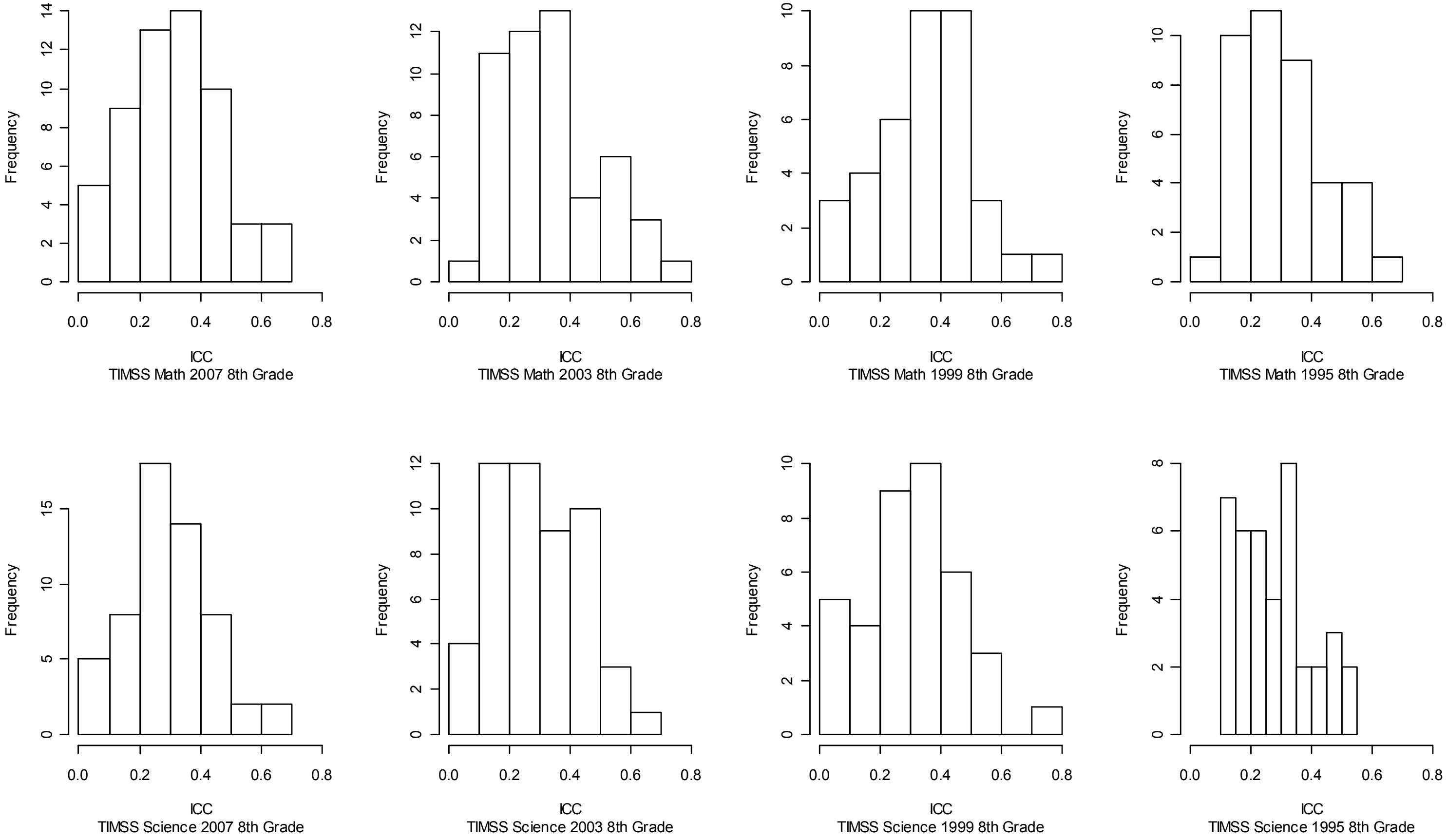
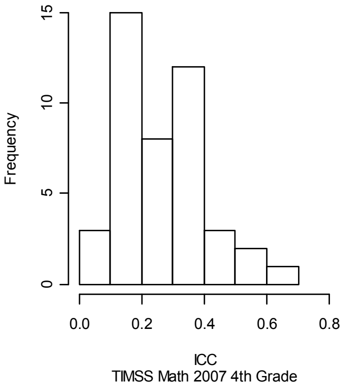


Figure 2. The Change in Average Two-Level Fully Unconditional Intra-Class Correlation Coefficients across Years

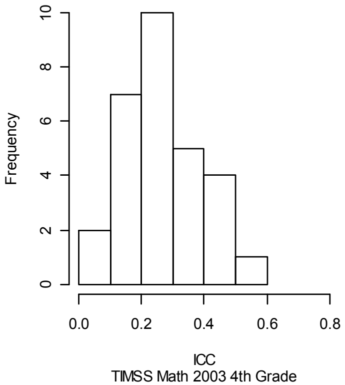
APPENDIX A

The Distribution of Two-Level Fully Unconditional Intra-Class Correlation Coefficients across Countries at Different Levels of Grade, Achievement Domain, and Study Year

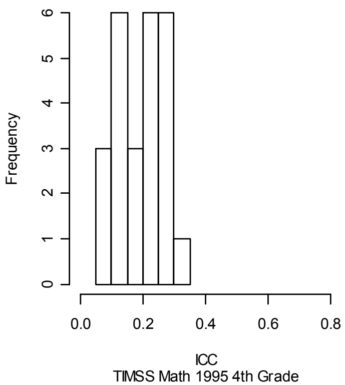

Frequency
ICC
0.0
0.2
0.4
0.6
0.8
TIMSS Math 2007 8th Grade
TIMSS Math 2003 8th Grade
TIMSS Math 1999 8th Grade
TIMSS Math 1995 8th Grade
TIMSS Science 2007 8th Grade
TIMSS Science 2003 8th Grade
TIMSS Science 1999 8th Grade
TIMSS Science 1995 8th Grade
0
2
4
6
8
10
12
14
5
15



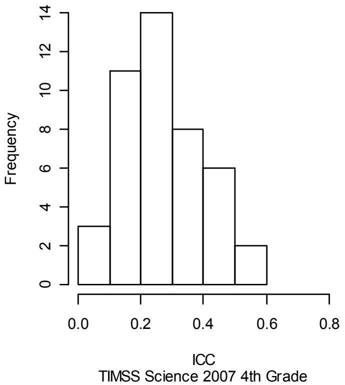
TIMSS Math 2007 4th Grade



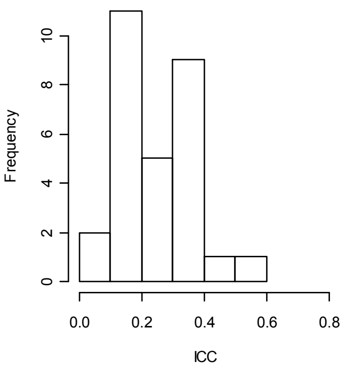
TIMSS Math 2003 4th Grade



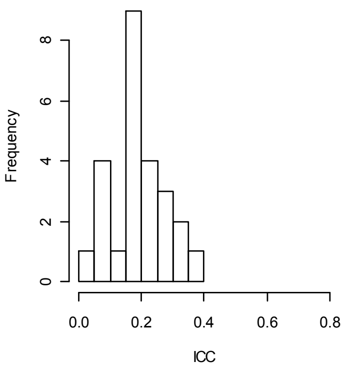
TIMSS Math 1995 4th Grade



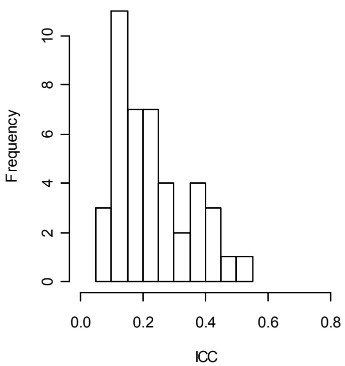
TIMSS Science 2007 4th Grade



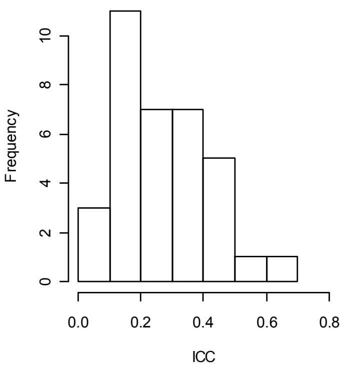
TIMSS Science 2003 4th Grade



TIMSS Science 1995 4th Grade



PIRLS Reading 2006 4th Grade



PIRLS Reading 2001 4th Grade

APPENDIX B

The Exact Intra-Class Correlation Coefficient for All Countries Estimated From Two-Level Fully Unconditional Model at Different Levels of Grade, Achievement Domain, and Study Year

Domain	Mathematics							Science							Reading	
Year	2007		2003		1999	1995		2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	8th	4th	8th	4th	8th	8th	4th	4th	4th
Country																
ADU	0.54	0.48	-	-	-	-	-	0.44	0.42	-	-	-	-	-	-	-
ARG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.46
ARM	0.22	0.34	0.18	0.30	-	-	-	0.30	0.42	0.27	0.33	-	-	-	-	-
AUS	0.54	0.30	0.48	0.29	0.55	0.37	0.26	0.48	0.28	0.37	0.25	0.40	0.32	0.22	-	-
AUT	-	0.13	-	-	-	0.34	0.25	-	0.11	-	-	-	0.31	0.24	0.14	-
BFL	-	-	0.67	0.16	0.39	0.40	-	-	-	0.57	0.14	0.26	0.28	-	0.14	-
BFR	-	-	-	-	-	0.38	-	-	-	-	-	-	0.31	-	0.20	-
BGR	0.41	-	0.33	-	0.49	-	-	0.39	-	0.33	-	0.39	-	-	0.36	0.39
BHR	0.22	-	0.18	-	-	-	-	0.21	-	0.11	-	-	-	-	-	-
BIH	0.17	-	-	-	-	-	-	0.19	-	-	-	-	-	-	-	-
BLZ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.42
BSQ	0.29	-	0.21	-	-	-	-	0.22	-	0.13	-	-	-	-	-	-
BWA	0.27	-	0.23	-	-	-	-	0.23	-	0.25	-	-	-	-	-	-
CAB	-	0.22	-	-	-	-	-	-	0.23	-	-	-	-	-	0.16	-
CAN	-	-	-	-	0.25	0.25	0.21	-	-	-	-	0.18	0.26	0.17	-	-
CBC	0.43	0.19	-	-	-	-	-	0.40	0.20	-	-	-	-	-	0.15	-
CHE	-	-	-	-	-	0.59	-	-	-	-	-	-	0.54	-	-	-
CHL	-	-	0.51	-	0.37	-	-	-	-	0.35	-	0.31	-	-	-	-
CNS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.14	-
COL	0.46	0.49	-	-	-	0.63	-	0.41	0.42	-	-	-	0.47	-	-	0.49
COT	0.30	0.23	0.13	0.20	-	-	-	0.31	0.23	0.12	0.18	-	-	-	0.13	0.17
CQU	0.42	0.17	0.38	0.14	-	-	-	0.36	0.14	0.26	0.11	-	-	-	0.15	0.18
CSK	-	-	-	-	-	-	0.14	-	-	-	-	-	-	0.15	-	-

Domain	Mathematics							Science							Reading	
Year	2007		2003		1999	1995		2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	8th	4th	8th	4th	8th	8th	4th	4th	4th
Country																
CYP	0.03	-	0.06	0.10	0.09	0.10	0.18	0.06	-	0.09	0.11	0.08	0.11	0.18	-	0.13
CZE	0.32	0.18	-	-	0.29	0.27	-	0.23	0.17	-	-	0.22	0.21	-	-	0.15
DEU	-	0.40	-	-	-	0.57	-	-	0.36	-	-	-	0.49	-	-	-
DNK	-	0.12	-	-	-	0.10	-	-	0.13	-	-	-	0.11	-	0.13	-
DZA	0.09	0.38	-	-	-	-	-	0.08	0.36	-	-	-	-	-	-	-
EGY	0.28	-	0.30	-	-	-	-	0.25	-	0.28	-	-	-	-	-	-
ENG	0.55	0.16	0.51	0.24	0.45	0.32	0.23	0.47	0.17	0.42	0.20	0.39	0.31	0.19	0.20	0.18
ESP	-	-	-	-	-	0.23	-	-	-	-	-	-	0.20	-	0.23	-
EST	-	-	0.22	-	-	-	-	-	-	0.16	-	-	-	-	-	-
FIN	-	-	-	-	0.14	-	-	-	-	-	-	0.10	-	-	-	-
FRA	-	-	-	-	-	0.25	-	-	-	-	-	-	0.19	-	0.15	0.16
GEO	0.31	0.39	-	-	-	-	-	0.23	0.37	-	-	-	-	-	0.39	-
GHA	0.42	-	0.33	-	-	-	-	0.44	-	0.32	-	-	-	-	-	-
GRC	-	-	-	-	-	0.18	0.25	-	-	-	-	-	0.25	0.28	-	0.24
HKG	0.65	0.31	0.58	0.24	0.57	0.56	0.21	0.57	0.28	0.46	0.19	0.41	0.43	0.16	0.27	0.28
HUN	0.30	0.30	0.32	0.25	0.30	0.20	0.16	0.26	0.26	0.25	0.21	0.24	0.18	0.16	0.25	0.23
IDN	0.45	-	0.53	-	0.44	-	-	0.43	-	0.46	-	0.41	-	-	0.38	-
IRL	-	-	-	-	-	0.35	0.15	-	-	-	-	-	0.31	0.17	-	-
IRN	0.32	0.37	0.31	0.29	0.25	0.28	-	0.31	0.37	0.24	0.25	0.26	0.30	-	0.41	0.40
ISL	-	-	-	-	-	0.12	0.07	-	-	-	-	-	0.13	0.08	0.08	0.09
ISR	0.33	-	0.33	-	0.36	0.27	0.12	0.31	-	0.25	-	0.37	0.23	0.16	0.42	0.39
ITA	0.23	0.33	0.27	0.32	0.35	0.27	-	0.26	0.31	0.25	0.32	0.30	0.20	-	0.28	0.20
JOR	0.31	-	0.26	-	0.24	-	-	0.28	-	0.21	-	0.22	-	-	-	-

Domain	Mathematics							Science							Reading	
Year	2007		2003		1999	1995		2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	8th	4th	8th	4th	8th	8th	4th	4th	4th
Country																
JPN	0.20	0.08	0.18	0.04	0.10	0.17	0.05	0.17	0.08	0.12	0.04	0.07	0.13	0.04	-	-
KAZ	-	0.56	-	-	-	-	-	-	0.48	-	-	-	-	-	-	-
KOR	0.09	-	0.13	-	0.10	0.11	0.13	0.06	-	0.08	-	0.07	0.11	0.08	-	-
KWT	0.20	0.24	-	-	-	0.18	0.19	0.25	0.25	-	-	-	0.25	0.18	0.21	0.32
LBN	0.43	-	0.42	-	-	-	-	0.52	-	0.40	-	-	-	-	-	-
LTU	0.18	0.15	0.20	0.23	0.34	0.21	-	0.15	0.11	0.15	0.15	0.25	0.18	-	0.13	0.22
LUX	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.17	-
LVA	-	0.20	0.21	0.25	0.22	0.21	0.25	-	0.23	0.18	0.25	0.19	0.17	0.31	0.22	0.19
MAR	0.37	0.62	0.14	0.44	0.07	-	-	0.35	0.58	0.16	0.34	0.06	-	-	0.52	0.65
MDA	-	-	0.34	0.38	0.41	-	-	-	-	0.32	0.34	0.31	-	-	0.22	0.38
MKD	-	-	0.38	-	0.34	-	-	-	-	0.37	-	0.34	-	-	0.46	0.44
MLT	0.63	-	-	-	-	-	-	0.61	-	-	-	-	-	-	-	-
MNG	0.27	0.31	-	-	-	-	-	0.19	0.26	-	-	-	-	-	-	-
MYS	0.61	-	0.53	-	0.59	-	-	0.61	-	0.48	-	0.49	-	-	-	-
NLD	-	0.18	0.71	0.18	0.70	0.54	0.13	-	0.18	0.61	0.15	0.54	0.44	0.15	0.17	0.20
NOR	0.09	0.12	0.13	0.09	-	0.14	0.07	0.08	0.14	0.09	0.09	-	0.14	0.05	0.10	0.08
NZL	-	0.21	0.43	0.30	0.49	0.35	0.28	-	0.22	0.42	0.34	0.40	0.36	0.31	0.31	0.29
OMN	0.21	-	-	-	-	-	-	0.21	-	-	-	-	-	-	-	-
PHL	-	-	0.61	0.43	0.48	-	-	-	-	0.56	0.38	0.53	-	-	-	-
POL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.09	-
PRT	-	-	-	-	-	0.22	0.20	-	-	-	-	-	0.17	0.24	-	-
PSE	0.19	-	0.17	-	-	-	-	0.18	-	0.14	-	-	-	-	-	-
QAT	0.25	0.19	-	-	-	-	-	0.40	0.32	-	-	-	-	-	0.15	-

Domain	Mathematics							Science							Reading	
Year	2007		2003		1999	1995		2007		2003		1999	1995		2006	2001
Grade	8th	4th	8th	4th	8th	8th	4th	8th	4th	8th	4th	8th	8th	4th	4th	4th
Country																
ROM	0.30	-	0.37	-	0.42	0.44	-	0.30	-	0.40	-	0.38	0.47	-	0.35	0.38
RUS	0.36	0.56	0.36	0.46	0.42	0.35	-	0.30	0.54	0.34	0.42	0.41	0.35	-	0.43	0.46
SAU	0.18	-	0.22	-	-	-	-	0.22	-	0.24	-	-	-	-	-	-
SCG	0.18	-	0.13	-	-	-	-	0.16	-	0.14	-	-	-	-	-	-
SCO	0.42	0.13	0.56	0.14	-	0.27	0.21	0.38	0.14	0.44	0.16	-	0.30	0.22	0.14	0.18
SGP	0.47	0.23	0.39	0.56	0.71	0.48	0.24	0.50	0.25	0.42	0.51	0.75	0.50	0.27	0.23	0.58
SLV	0.32	0.39	-	-	-	-	-	0.35	0.41	-	-	-	-	-	-	-
SVK	-	0.36	0.35	-	0.32	0.16	-	-	0.38	0.30	-	0.26	0.22	-	0.29	0.25
SVN	0.09	0.07	0.12	0.13	0.13	0.12	0.10	0.09	0.09	0.08	0.12	0.11	0.10	0.09	0.10	0.09
SWE	0.11	0.16	0.25	-	-	0.30	-	0.13	0.18	0.22	-	-	0.31	-	0.10	0.14
SYR	0.40	-	0.23	-	-	-	-	0.28	-	0.19	-	-	-	-	-	-
THA	0.39	-	-	-	0.45	0.31	0.34	0.37	-	-	-	0.42	0.25	0.35	-	-
TTO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.36	-
TUN	0.16	0.28	0.18	0.38	0.17	-	-	0.11	0.27	0.12	0.35	0.15	-	-	-	-
TUR	0.35	-	-	-	0.30	-	-	0.31	-	-	-	0.29	-	-	-	0.30
TWN	0.28	0.10	0.39	0.27	0.22	-	-	0.24	0.08	0.34	0.13	0.20	-	-	0.15	-
UIN	-	-	0.29	0.18	-	-	-	-	-	0.28	0.20	-	-	-	-	-
UKR	0.24	0.21	-	-	-	-	-	0.22	0.15	-	-	-	-	-	-	-
UMA	0.34	0.15	-	-	-	-	-	0.36	0.22	-	-	-	-	-	-	-
UMN	0.27	0.19	-	-	-	-	-	0.24	0.23	-	-	-	-	-	-	-
USA	0.32	0.29	0.42	0.33	0.34	0.41	0.25	0.31	0.31	0.43	0.37	0.35	0.40	0.30	0.23	0.27
YEM	-	0.45	-	0.42	-	-	-	-	0.46	-	0.39	-	-	-	-	-
ZAF	-	-	0.61	-	0.48	-	-	-	-	0.57	-	0.57	-	-	-	-