P25: INTRACLASS CORRELATION AND VARIANCE COMPONENTS AS POPULATION ATTRIBUTES AND MEASURES OF SAMPLING EFFICIENCY IN PIRLS 2001

Pierre Foy, IEA Data Processing Center, Hamburg, Germany

Keywords: Intraclass Correlation, Variance Decomposition, Sample Design, Stratification

Abstract

The Intraclass Correlation is a measure of homogeneity among analytical units. In the context of PIRLS 2001, the intraclass correlation measures the homogeneity of reading achievement between schools in a national educational system. It allows the researcher to address policy concerns related to equity and disparity of learning opportunities. The derivation of variance components, through variance decomposition, allows the researcher to probe deeper into the various sources of variance in reading achievement. Specifically, we look at students, classes and schools as sources of variance. We further break down the school variance component into a school stratification component to investigate disparity of reading achievement between groupings of schools, as defined by the stratification variables applied in national sample designs. This latter component also allows us to measure the efficiency of the school-level stratifications, as applied nationally, in reducing the standard errors of the survey estimates.

Introduction

This paper presents two related concepts, as applied to the PIRLS 2001 data. The intraclass correlation is a measure of homogeneity among units of analysis, and thus provides a summary glimpse into the variance structure of the data. Variance components are derived from variance decomposition and offer a more detailed perspective into the variance structure. We will describe in detail these two concepts; demonstrate how they are computed, with results taken from the PIRLS 2001 data, and show how the results can be interpreted, both as population attributes and as measures of sampling efficiency.

Intraclass Correlation

Definition

The Intraclass Correlation is an important population attribute used by sample survey statisticians in deriving efficient sample designs and sample sizes for hierarchical populations. The PIRLS data are hierarchical in nature, with students within classes, and classes within schools. At minimum, we can derive two variance components: the between-school variance and the within-school variance. The intraclass correlation simply expresses the between-school variance as the proportion of the sum of the two variance components, as described in the following equation:

$$IC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \quad (1)$$

where σ_B^2 is the between-school variance, and σ_W^2 is the within-school variance.

A low intraclass correlation, say less than 0.25, indicates relatively small between-school variations. In other words, schools tend to perform at comparable levels. As the intraclass correlation increases, beyond 0.25, then schools perform with ever-increasing variations; some schools achieving very high levels of performance and others very low levels of performance.

For the sampling statistician, a school system with a low intraclass correlation requires a sample design that focuses more on the within-school component. Thus he will devise a sample design that samples fewer schools, but more students within schools. As the intraclass correlation increases, the focus shifts to sampling more schools, and perhaps fewer students within schools.

As a population attribute, the intraclass correlation offers a measure of equity, or disparity, of learning opportunity. Systems with a low intraclass correlation have achieved a measure of equity whereby all schools perform at roughly equivalent levels. Systems with a high intraclass correlation demonstrate disparity of learning opportunity whereby some schools perform well, yet others in the same system perform poorly.

Estimation

The Intraclass Correlation is a simple case of variance decomposition. It is derived from a single-level analysis of variance, as presented in figure 1 (Neter & Wasserman, p. 442). The ANOVA table presents the general case with unequal student sample sizes within schools, and is computed using sampling weights[1].

Figure 1: ANOVA Table for Computing the Intraclass Correlation

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	E(MS)
Between schools	$SS_B = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2$	$n-1$	$MS_B = \frac{SS_B}{n-1}$	$\sigma_W^2 + n'\,\sigma_B^2$
Within schools	$SS_W = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(y_{ij} - \bar{y}_{i.})^2$	$\sum_{i=1}^{n}(n_i - 1)$	$MS_W = \frac{SS_W}{\sum_{i=1}^{n}(n_i - 1)}$	σ_W^2
Total	$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(y_{ij} - \bar{y}_{..})^2$	$\sum_{i=1}^{n} n_i - 1$		

The quantity n' is estimated as follows (Neter & Wasserman, p. 528):

$$n' = \frac{1}{n-1}\left[\sum_{i=1}^{n} n_i - \sum_{i=1}^{n} n_i^2 \Big/ \sum_{i=1}^{n} n_i\right] \quad \textbf{(2)}$$

where n is the number of sampled schools, and n_i is the number of sampled students in school "i". This quantity can be interpreted as the average n_i in the case of an unbalanced ANOVA. If all $n_i = k$, as is the case in a balanced ANOVA, then $n' = k$.

The intraclass correlation can thus be estimated from the mean squares in the ANOVA table and using equation (1), as follows:

$$IC = \frac{MS_B - MS_W}{MS_B + (n'-1)\,MS_W} \quad \textbf{(3)}$$

where MS_B and MS_W are defined in the ANOVA table of figure 1. In an unbalanced ANOVA, the quantity n' must be used. In a balanced ANOVA, n' can be replaced with the constant student sample size k in each school.

For the analysis of variance model to work properly, some basic assumptions on the nature of the data being analysed must hold:

1. *Two levels*: The data come from a hierarchical structure with two levels, namely schools and students. This may seem obvious, but needs to be stated since in some respects we may be applying a simpler model to a more complex data structure.

[1] The sampling weights used are the "house weights", as described in the "PIRLS 2001 User Guide for the International Database" (Gonzalez & Kennedy, 2003).

2. *Equal probability sampling*: Units at each level must be selected with equal probabilities. In this case, schools should be sampled with equal probabilities, and students within schools should be selected with equal probabilities, although these student probabilities need not be the same between schools.
3. *Equal variance*: We require equal variance at all levels. In this case, the student within school variance should be equal for each school.
4. *Random effects*: Units at each level must consist of a random sample from a larger population. Thus schools must be sampled from a larger population and students must be sampled from a larger population of students within each school.

As we will see later in the interpretation of actual results, these assumptions do not always hold. And we might therefore want to reconsider our approach to this problem. The assumption of random effects is perhaps the least restrictive since we usually have school and student samples taken from larger populations. All of these assumptions must hold if we are to make valid statistical tests from the resulting data. The violation of any of these assumptions, however, does not prevent us from deriving estimates for descriptive purposes.

Application

We have applied equation (3) to the PIRLS 2001 data. The data were computed using the procedure NESTED from SAS, version 8. The PIRLS study was carried out in 2001 and administered a reading test at the 4^{th} grade in 35 countries. The results are given in Table 1. The table gives intraclass correlations for the overall reading score, as well as for its two components: *reading for literary experience* and *reading to acquire and use information*. All calculations are based on the first plausible value of each score. The countries are ranked in ascending order of the overall reading intraclass correlation.

The countries have been arbitrarily divided into three groups. The first group consists of countries with a low intraclass correlation, the cut-off being set at 0.25. This means that, for these countries, less than 25% of the total variance is between schools; for Iceland, the between-school variance for overall reading represents only 8% of the total variance.

At the other extreme of the intraclass correlation spectrum, we find a group of countries with relatively large intraclass correlations. This group is arbitrarily defined as those with intraclass correlations grater than 0.35, i.e., the between-school variance represents more than 35% of the total variance; for Singapore, the between-school variance accounts for as much as 56% of the total variance.

Between these two groups, we find a third one, where intraclass correlations lie between 0.25 and 0.33.

The break points used to define the three groups of countries are to some extent arbitrary, we can nonetheless refer to countries having either a low intraclass correlation, a high intraclass correlation, or somewhere in between. Countries with a low intraclass correlation have achieved some measure of equity in learning achievement. All countries should be interested in discovering the context for achieving this result, using data from this group of countries. In all three groups, we should be interested in better understanding the source of the between-school variance, albeit to varying degrees. This line of enquiry takes on more relevance as the intraclass correlation increases.

Although countries with a low intraclass correlation might provide contextual information to describe how they have reached some state of equity in learning opportunity, they might also want to check this information as confirmation of their efforts. For example, it might be interesting to discover that a particular country still manages equity without necessarily having a perfectly fair allocation of certain resources at the school or classroom level.

It must be said that having equated a low intraclass correlation with equity in educational achievement, equity being perceived as a positive outcome, this does not necessarily entail high educational achievement. In fact, a simple regression between national achievement scores and their

intraclass correlation yields a small negative slope (-0.001, yet statistically significant). One need only look at the four countries with the lowest intraclass correlations (Iceland, Slovenia, Norway and Cyprus) to realise this. This is further confounded by the fact that these countries are among those with the lowest student mean ages. This relationship needs further study; investigating the relationship between age and student achievement across countries in a grade-based study. We can speculate that the intraclass correlation might increase as the students progress through the grades, but this has not been demonstrated in past IEA studies (i.e., TIMSS), where the intraclass correlation at the 8th grade, for both mathematics and science, remains at comparable levels for these four countries in particular.

Table 1: Intraclass Correlations for PIRLS 2001

Country	Overall Reading	Reading for Literary Experience	Reading to Acquire and Use Information
Iceland	0.086	0.092	0.095
Slovenia	0.090	0.093	0.069
Cyprus	0.101	0.092	0.101
Norway	0.104	0.075	0.098
Sweden	0.138	0.149	0.125
Germany	0.146	0.170	0.134
Czech Republic	0.162	0.142	0.154
France	0.167	0.159	0.170
Canada (O,Q)	0.175	0.187	0.173
England	0.183	0.167	0.196
Scotland	0.185	0.160	0.192
Netherlands	0.195	0.180	0.196
Italy	0.200	0.200	0.217
Latvia	0.218	0.206	0.189
Lithuania	0.220	0.216	0.220
Hungary	0.223	0.205	0.207
Greece	0.231	0.224	0.225
Slovak Republic	0.256	0.223	0.227
New Zealand	0.258	0.260	0.259
Turkey	0.274	0.244	0.292
United States	0.275	0.248	0.301
Hong Kong (SAR)	0.297	0.273	0.287
Kuwait	0.335	0.297	0.306
Bulgaria	0.348	0.347	0.314
Belize	0.348	0.367	0.315
Romania	0.363	0.348	0.337
Iran	0.392	0.415	0.361
Moldova	0.398	0.356	0.382
Israel	0.417	0.424	0.353
Argentina	0.421	0.411	0.374
Macedonia	0.427	0.443	0.449
Russian Federation	0.454	0.373	0.453
Colombia	0.475	0.443	0.445
Morocco	0.559	0.502	0.558
Singapore	0.587	0.572	0.585

It is also important to mention that countries with the larger intraclass correlations tend to be among the low achievers. Consequently, countries with a high intraclass correlation should investigate the causes of their large between-school variances, as potential means for improving their overall performance levels.

From the sampling perspective, the magnitude of the intraclass correlation is an indicator of the intrinsic inefficiencies in drawing samples from a hierarchical population. As the intraclass correlation increases, sampling becomes less efficient and requires more efficient sample designs, larger sample sizes, or both. Therefore, populations with low intraclass correlations place fewer demands on the sampling statistician in developing a suitable sample design. Populations with large intraclass correlations require additional knowledge of the nature of the disparities between schools, information that can be used in defining strata in such a way as to minimise the impact of the large intraclass correlation. The final weapon available is simply to increase the sample size, although, once again, additional knowledge of the true nature of disparity might require either a larger sample of schools, or a larger sample of students or classrooms within schools, depending on which is the major source of variance.

Interpretation

Because the PIRLS 2001 data are derived from complex, stratified, multi-stage sample designs, proper interpretation of the calculated intraclass correlations must take these sample design features in consideration. This is particularly important to the extent that the data derived from these complex sample designs allow us to meet, or violate, the underlying assumptions supporting the analysis of variance model, as presented earlier.

PIRLS participants first defined strata, then selected schools, classrooms and finally students. Most participants sampled only one classroom per school, whereas a few others sampled more than one. This is an important distinction to make since classrooms within schools can be another source of variance. Given the assumptions behind the simplified analysis of variance model in Figure 1, this within-classroom variance component will either be confounded with the between-school variance component for countries that sampled only one classroom per school, or confounded with the within-school variance component for countries that sampled more than one classroom per school. Therefore, we have some imbalance in the interpretation of the intraclass correlations in Table 1, since countries that sampled only one classroom per school will tend to have a larger intraclass correlation than if they had sampled more than one classroom per school. This of course depends on the magnitude of the within-classroom variance component. If this component is small or non-existent, then sampling more than one classroom per school makes little difference. But if a country sampled only one classroom per school, it is not possible to establish the magnitude of the between-classroom variance component.

The following countries systematically sampled more than one classroom per school: Colombia, France, Germany, Iceland, Iran, Kuwait, Netherlands, Norway and Sweden. For these countries, the calculated intraclass correlations are better reflections of their true between-school variance components. The intraclass correlations for all other countries will tend to be an over-estimation of the true between-school variance components, depending on the magnitude of their unknown between-classroom variance components.

In a scenario whereby only one classroom per school is sampled, we can claim to have a sample of classrooms, as opposed to a sample of schools, from the whole population. If we were to redefine the structure of the ANOVA table in Figure 1 in this context, we would label the sources of variation as “between classrooms” and “within classrooms”. Thus the intraclass correlation becomes more a measure of the disparity between classrooms in the population than between schools. We make here a distinction between “between classrooms in the population” and “between classrooms within schools”, the former expected to be greater or equal to the latter. Thus if we were to calculate the intraclass correlations considering classrooms as the first level, rather than schools, we would obtain more comparable results between countries since we would consider all samples on the same footing, that is to say as samples of classrooms as opposed to samples of schools.

The data in Table 2 allow us to make this comparison. The “school level” column considers the PIRLS samples as samples of schools and is identical to the overall reading column of Table 1. The “classroom level” column considers the PIRLS samples as samples of classrooms, rather than schools, and thus we would expect the numbers in this column to be greater for countries that sampled more than one classroom per school. This is indeed what we observe in the “increase” column, which describes the increase in the classroom column, relative to the school column. This increase gives us a

glimpse of the presence of a between-classroom within school variance component, without explicitly measuring it.

Table 2: Comparing Intraclass Correlations Between Schools and Between Classrooms

Country	Overall Reading		
	School Level	Classroom Level	Increase
Iceland	0.086	0.115	33.6%
Slovenia	0.090	0.092	—
Cyprus	0.101	0.101	—
Norway	0.104	0.110	5.0%
Sweden	0.138	0.174	25.9%
Germany	0.146	0.167	14.3%
Czech Republic	0.162	0.162	—
France	0.167	0.206	23.3%
Canada (O,Q)	0.175	0.182	—
England	0.183	0.183	—
Scotland	0.185	0.195	—
Netherlands	0.195	0.204	4.3%
Italy	0.200	0.200	—
Latvia	0.218	0.219	—
Lithuania	0.220	0.220	—
Hungary	0.223	0.223	—
Greece	0.231	0.231	—
Slovak Republic	0.256	0.277	8.4%
New Zealand	0.258	0.270	—
Turkey	0.274	0.274	—
United States	0.275	0.284	—
Hong Kong (SAR)	0.297	0.297	—
Kuwait	0.335	0.374	11.4%
Bulgaria	0.348	0.348	—
Belize	0.348	0.376	7.9%
Romania	0.363	0.377	3.8%
Iran	0.392	0.399	1.8%
Moldova	0.398	0.398	—
Israel	0.417	0.417	—
Argentina	0.421	0.421	—
Macedonia	0.427	0.448	4.9%
Russian Federation	0.454	0.454	—
Colombia	0.475	0.484	2.0%
Morocco	0.559	0.559	—
Singapore	0.587	0.587	—

The reader should note that there are more countries here with positive increases than the countries listed earlier that systematically sampled more than one classroom per school. This is due to some countries having sampled more than one classroom per school, but in a limited way, albeit enough to be reliably estimated in this table. It is interesting to note that the intraclass correlation in Iceland increases dramatically, suggesting a possibly large classroom variance component. This is in stark contrast to Norway, where we observe a much smaller increase.

We should point out the limitations of sampling only one classroom per school in being able to make a distinction between school and classroom variances. Taking Cyprus as an example, had we sampled more than one classroom per school, we would expect to estimate the same value in the "classroom level" column, but a value less or equal in the "school level" column, the magnitude of the difference dependent on the magnitude of the between-classroom within school variance component.

The interplay between the between-school and between-classroom variance components raises other potential shortcomings in the proper interpretation of the intraclass correlations. Our first assumption behind the analysis of variance used to derive the intraclass correlation is that the data are derived from a two-level model, namely schools and the students within. This assumption does not quite hold in the context of PIRLS where the sample design is at minimum a three-level model: schools with classrooms and students. Having said that, we are still permitted to view the data through a two-level model, either schools and students, or classrooms and students, provided the various sources of variance remain uncorrelated and we do not violate the other underlying assumptions.

Our second assumption is that all units within a level are selected with equal probabilities. This actually holds for classrooms and students, but does not hold for schools, since they are selected using a stratified PPS (probabilities proportional to size) sample design. If the between-school variance component is constant, regardless of school size, then we can expect minimal disturbance from the PPS sample selection. This assumption of equal probabilities is more readily met if we consider classrooms instead of schools as the first level, since the classroom selection probabilities have almost equal probabilities by design.

Our third assumption is one of equal variance at each level. Specifically, the between-classroom variance should be equal in all schools. Intuitively, we should question whether this assumption holds or not. Let us first consider the student level variance within schools. Whether a school has one classroom of students, or ten classrooms of students, let us for the sake of argument assume that we have equal student-level variance. The between-variance component arises from the partitioning of students into classes, and this process apportions some of the student-level variance to a between-classroom component and the remainder to a within-classroom component. Our contention is that the magnitude of the between-classroom component can be highly dependent on the number of classrooms to be formed in a given school. The total within-school variance may be the same in all schools, but there can be more classroom differentiation in a school with ten classes than in a school with only two classrooms. Thus we would expect a larger between-classroom variance component in larger schools than in smaller schools, thereby invalidating our assumption of equal variance at the classroom-level. This phenomenon remains to be tested and validated.

We formulated a fourth assumption, one of random effects at all levels. This is clearly validated given that we have sampled units at all levels: schools, classrooms and students. Only for a handful of countries did we select all eligible schools, and thus this assumption might not hold. For these countries, we could infer a super-population model, whereby the relatively small numbers of schools in their samples are conceptually samples of their respective super-populations.

This discussion leads us to consider a more elaborate analysis of variance model, better suited for the complex structure of the PIRLS data. This brings us then to consider a four-level model, which would allow us to explore in greater detail, the many variance components, some of which we have alluded to already, present in the PIRLS data.

Variance Decomposition

Definition

Recognizing that the PIRLS data are hierarchical in nature – students within classrooms, classrooms within schools and schools within sampling strata – we can explore the variance structure of these data using a four-level analysis of variance model. This model is presented in Figure 2 (Bortz, p. 532), and is an extension of the simpler model presented in Figure 1. Thus all terms in the figure are analogous and hopefully obvious, except perhaps for the derivation of the quantities n'' and n'''. They are derived in an analogous manner as n' and are meant to reflect the average student sample size at the school level and at the stratum level respectively.

The underlying assumptions in the four-level analysis of variance model are similar to the ones discussed earlier for the two-level model:

1. Four *levels*: We now consider that the data come from a hierarchical structure with four levels, namely strata, schools, classrooms and students.

2. *Equal probability sampling*: Units at each level must be selected with equal probabilities. We have a difficulty here since strata are not actually sampled, but this is more relevant under our fourth assumption.
3. *Equal variance*: We again require equal variance at all levels. We have already discussed a potential difficulty in this regard concerning the between-classroom variance component.
4. *Random effects*: Units at each level must consist of a random sample from a larger population. This clearly does not hold for the strata, and it is difficult to conceive of a super-population model suitable for this level.

Figure 2: ANOVA Table for Variance Decomposition

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	E(MS)
Between strata	$SS_{str} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} w_{hijk}(\bar{y}_{h...} - \bar{y}_{....})^2$	$H-1$	$MS_{str} = \frac{SS_{str}}{H-1}$	$\sigma^2_{std} + n'\sigma^2_{cls} + n''\sigma^2_{sch} + n'''\sigma^2_{str}$
Between schools within strata	$SS_{sch} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} w_{hijk}(\bar{y}_{hi..} - \bar{y}_{h...})^2$	$\sum_{h=1}^{H}(n_h - 1)$	$MS_{sch} = \frac{SS_{sch}}{\sum_{h=1}^{H}(n_h - 1)}$	$\sigma^2_{std} + n'\sigma^2_{cls} + n''\sigma^2_{sch}$
Between classrooms within schools	$SS_{cls} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} w_{hijk}(\bar{y}_{hij.} - \bar{y}_{hi..})^2$	$\sum_{h=1}^{H}\sum_{i=1}^{n_h}(n_{hi} - 1)$	$MS_{cls} = \frac{SS_{cls}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}(n_{hi} - 1)}$	$\sigma^2_{std} + n'\sigma^2_{cls}$
Between students within classrooms	$SS_{std} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} w_{hijk}(y_{hijk} - \bar{y}_{hij.})^2$	$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}(n_{hij} - 1)$	$MS_{std} = \frac{SS_{std}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}(n_{hij} - 1)}$	σ^2_{std}
Total	$SS_{tot} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}}\sum_{k=1}^{n_{hij}} w_{hijk}(y_{hijk} - \bar{y}_{....})^2$	$\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} n_{hij} - 1$		

From the ANOVA table, we can estimate four variance components:

- A between-strata variance, σ^2_{str}
- A between-schools within strata variance, σ^2_{sch}
- A between-classrooms within schools variance, σ^2_{cls}
- A between-students within classrooms variance, σ^2_{std}

We can estimate the total variance as the sum of these four variance components, and then express each variance component as a percentage of this total variance. This is analogous to the definition of the intraclass correlation.

Application

Using the PIRLS 2001 data, we estimated all four variance components for all participating countries. This was done only with the overall reading score, using its first plausible value, and the "house weights". The estimates are presented in Table 3. The *Mean Square Error* is an estimate of the total variance, and the variance components, labelled *STR*, *SCH*, *CLS* and *STD* respectively, are expressed as percentages of the total variance. Countries are presented in alphabetical order.

At first glance, we will notice that the classroom component is estimated as zero (0) for most countries. This simply reflects the fact that these countries sampled only one classroom per school, or more than one classroom in too few schools, thereby making the estimation of this variance component impossible, or unreliable.

These data are also presented graphically in Figures 3 and 4. The pie charts display the relative magnitude of the four variance components. The total area of each pie chart is proportional to each country's mean square error, thus graphically displaying the prominence of national total variances.

For example, since Morocco has the largest total variance, it also has the largest chart. Iceland having roughly half the total variance of Morocco, it has a chart whose area is half that of Morocco.

Table 2: Variance Decomposition Data for PIRLS 2001

Country	Mean Square Error	Variance Components			
		STD	CLS	SCH	STR
Argentina	9 220	57.3%	—	23.2%	19.5%
Belize	11 324	60.9%	15.0%	9.4%	14.7%
Bulgaria	6 876	49.3%	—	22.0%	28.7%
Canada (O,Q)	5 071	82.1%	—	15.6%	2.3%
Colombia	6 518	46.7%	3.0%	30.3%	20.1%
Cyprus	6 748	89.5%	—	7.9%	2.6%
Czech Republic	4 091	83.8%	—	16.2%	0.0%
England	7 272	81.3%	—	12.7%	6.0%
France	4 967	75.4%	10.2%	3.7%	10.8%
Germany	4 467	83.2%	4.4%	10.6%	1.8%
Greece	5 282	76.3%	—	16.8%	7.0%
Hong Kong (SAR)	4 017	70.2%	—	24.0%	5.8%
Hungary	4 260	77.6%	—	17.9%	4.5%
Iceland	5 690	88.4%	5.6%	5.6%	0.5%
Iran	8 513	54.3%	1.8%	28.1%	15.8%
Israel	8 849	55.0%	—	8.1%	36.8%
Italy	5 043	79.8%	—	15.7%	4.4%
Kuwait	8 006	61.9%	7.7%	19.4%	11.0%
Latvia	3 686	77.6%	—	15.7%	6.8%
Lithuania	4 094	78.0%	—	22.0%	0.0%
Macedonia	10 452	50.8%	16.0%	2.1%	31.2%
Moldova	5 585	59.5%	—	29.2%	11.3%
Morocco	12 911	43.2%	—	56.8%	0.0%
Netherlands	3 309	79.3%	3.1%	14.5%	3.1%
New Zealand	8 734	71.9%	—	10.2%	17.9%
Norway	6 742	88.8%	1.5%	7.8%	1.9%
Romania	7 922	61.8%	8.3%	21.8%	8.0%
Russian Federation	4 601	54.4%	—	27.7%	17.9%
Scotland	7 151	81.3%	—	18.7%	0.0%
Singapore	8 284	41.3%	—	58.7%	0.0%
Slovak Republic	4 902	72.2%	13.3%	14.5%	0.0%
Slovenia	5 170	90.9%	—	8.3%	0.8%
Sweden	4 289	82.0%	6.4%	11.5%	0.0%
Turkey	7 506	72.4%	—	21.8%	5.8%
United States of America	7 019	71.8%	—	25.5%	2.7%

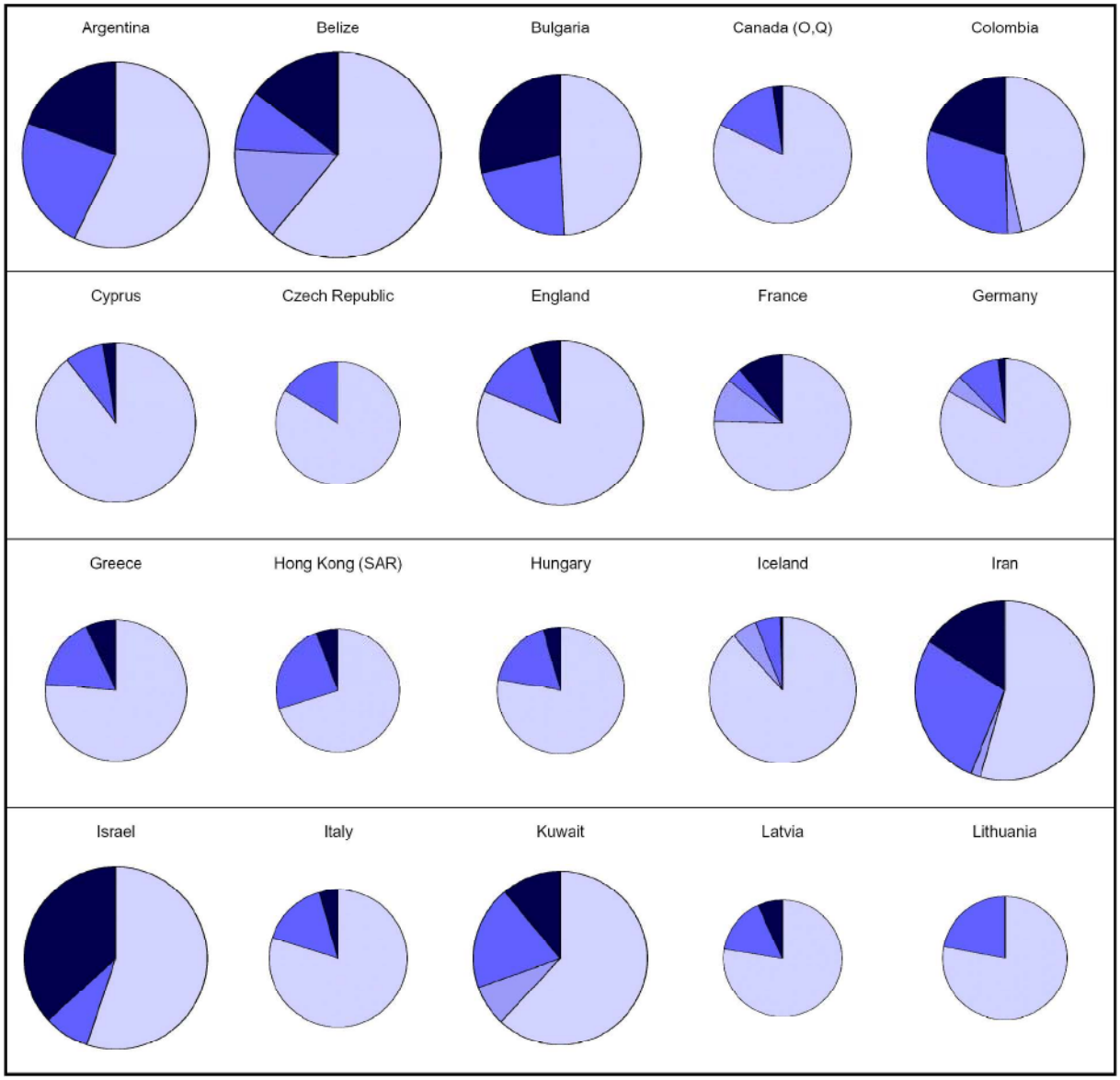
Interpretation

In theory, we should be able to observe the intraclass correlations from these data. We say in theory because this is not always the case. It is unclear at this time why this is so, but we assume it is related to the unbalanced nature of these ANOVAs. A balanced ANOVA has equal sample sizes at all levels and all the desired properties for properly estimating all of its parameters. The PIRLS data are not at all balanced, especially regarding the school sample sizes within strata, and to lesser degrees regarding classrooms within schools and students within classrooms. Another possible explanation for this less than perfect fit could be related to the potentially poor estimation of the classroom variance component. Earlier we discussed the possibility that it may not be appropriate to assume equal variance at the classroom level. This may be manifesting itself here, and definitely warrants further investigation.

Argentina is an example where we can readily observe the intraclass correlation as was estimated in Table 1. Recall that it was 0.421, meaning that the between-school variance represents 42.1% of the total variance. This between-school variance is now divided into two components: the

stratum and school components. Summing up these two components, we obtain a comparable estimate of 42.7% of total variance explained.

Figure 3: Variance Decomposition for PIRLS 2001 – Part 1

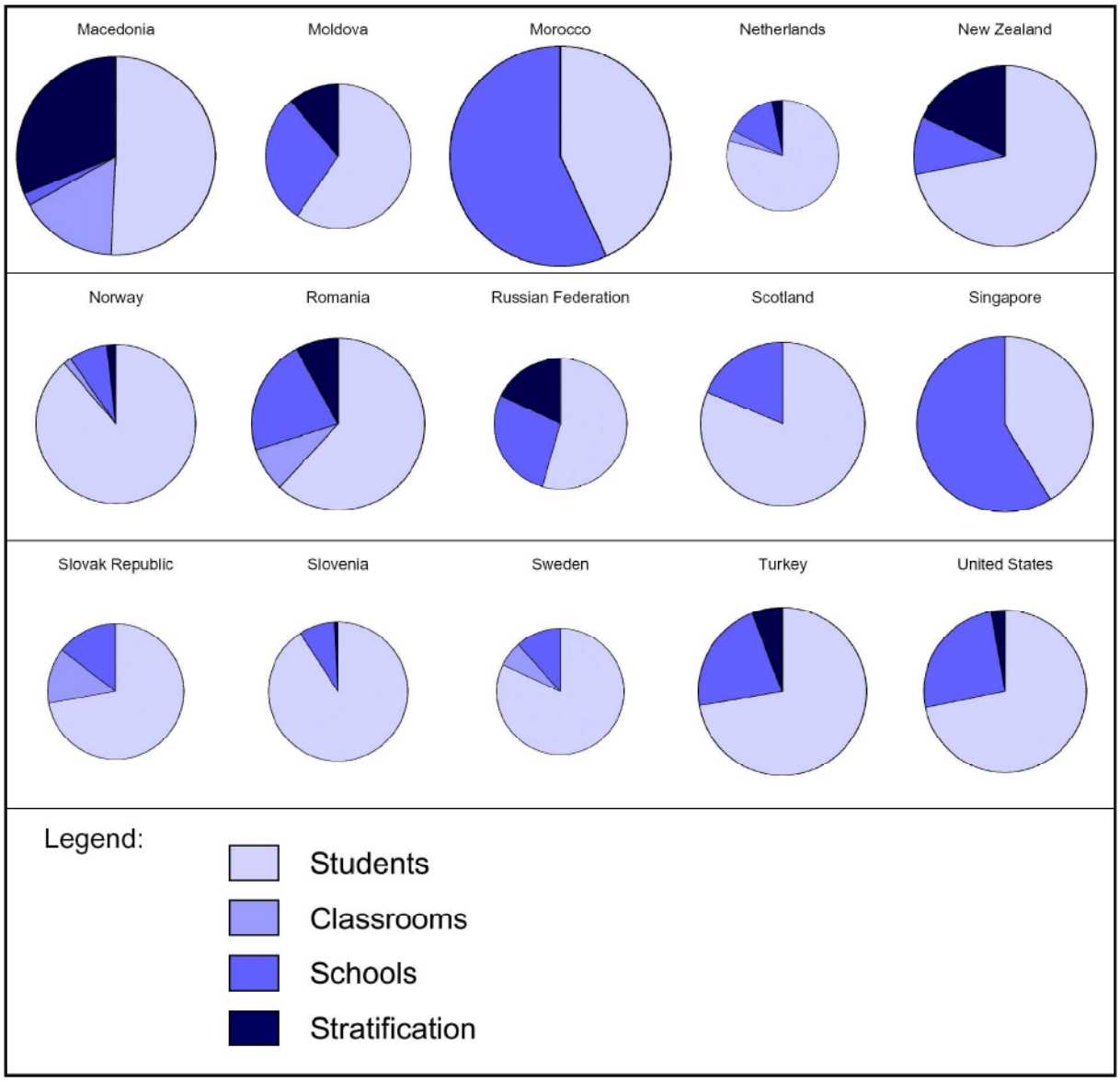


If we take Bulgaria as an example, its intraclass correlation was estimated at 0.348 in Table 1. The sum of its stratum and school components yields 50.7%. This rather large discrepancy cannot be explained and warrants further study. It is, however, the only country among those without a classroom component to display such a large discrepancy. All other countries without a classroom component behave much like Argentina, although with some varying minor discrepancies.

The situation is less obvious for those countries with a non-zero classroom variance component. If we were to sum up their stratum, school and classroom components, we would generally come close to their intraclass correlations, as estimated in the "classroom level" column of Table 2. This would tend to confirm our earlier inferences on the nature of the confounding of classroom variance when sampling only one classroom per schools versus sampling more than one classroom per school. We can then infer that whatever classroom variance may exist is currently confounded in the school variance for those countries that sampled only one classroom per school.

We must note the unusual results for the Czech Republic and the Slovak Republic. Although both countries did define sampling strata, we were not able to reliably estimate the stratum variance component. From closer examination of their data, sample sizes within strata appear very unbalanced and sparse, thus perhaps leading to the unreliability.

Figure 4: Variance Decomposition for PIRLS 2001 – Part 2



Stratification Efficiency

Despite the many caveats expressed regarding the relevance of the results derived from this multi-level unbalanced analysis of variance model, we can still glean many interesting and useful findings. From the sampling perspective, our primary interest is to explain as much as possible of the school variance component since this is the level which primarily dictates the ultimate reliability of the survey estimates derived from these data. Our purpose is to define school strata that will cluster schools according to their achievement levels. Countries did indeed define school strata, with varying levels of detail, based mostly on readily available administrative data sources. Our objective then in defining a stratum variance component in this variance decomposition was to evaluate the efficiency of these national stratification strategies. As we can see by the widely varying magnitudes of the

stratum components among participating countries, the efficiency of national stratification strategies varies greatly.

Efficient sampling strategies are evident in many countries, with Argentina and Colombia as obvious examples. In Argentina, 19.5% of the total variance is explained by differences between strata; a regional stratification contributing greatly to this efficiency. On the substantive side, this finding can also be useful since we can now acknowledge important regional differences in educational achievement. This has obvious policy implications, in setting objectives and mechanisms towards reducing regional disparities in the delivery of learning opportunities.

Regional stratification is a general trend in setting efficient sampling strategies among countries. Most of the countries with large intraclass correlations tend to highlight regional disparities (Argentina, Belize, Kuwait, Moldova, Morocco, Romania, Russian Federation & Turkey); at minimum we can find disparities between urban and rural parts (Colombia, Macedonia & Romania). School type, namely, public and private schools, can also be indicators of disparities in achievement (Argentina, Colombia & Iran). Language has also been used efficiently to stratify (Macedonia, Moldova & Turkey). Kuwait implemented a gender stratification that proved to be efficient.

The use of efficient stratification strategies does not benefit only countries with high intraclass correlations. Countries with low intraclass correlations also produced efficient implementations of similar stratification strategies, although with lesser impact since there was generally less between-school variance to apportion to a stratification variance component.

At the other end of the spectrum, we can readily find some countries with inefficient, or inadequate, stratification strategies. They can be identified by their small stratum variance component, such as Cyprus, Iceland, Lithuania, Morocco, the Netherlands, Scotland, Singapore, Slovenia, Sweden and Turkey. Those with low intraclass correlations can be granted some form of dispensation in this regard, considering there is less school variance to be apportioned to a stratum variance component. But some consideration of alternate stratification strategies should not be discounted.

Those countries with large intraclass correlations can ill afford inefficient stratification strategies since the alternative is to consider larger sample sizes, generally regarded as a burden on resources. Morocco and Singapore have the largest school variance components (43.2% & 41.3% respectively). Both should be encouraged to consider alternative stratification strategies. Preliminary investigations lead us to believe Morocco might want to consider school type in its stratification. Singapore did not implement any type of stratification, which seems plausible considering that nearly all schools are selected in the sample. Nonetheless, some from of stratification should perhaps be considered in the future. For TIMSS 2003, Singapore has chosen to sample two classrooms per school at the 8th grade, since they expect to have a large classroom component at that grade level. If this proves successful, then it could also be considered for the 4th grade in PIRLS.

The use of efficient stratification strategies is imperative for countries that have large standard errors in their published results (Martin & Mullis, p. 26). In PIRLS, we would include Argentina, Morocco and Singapore as candidates, since their standard errors are greater than the targeted 5 points. We have already discussed Morocco and Singapore. Since Argentina appears to have implemented a sound stratification strategy, the only other recourse would appear to be an increase in sample size.

Conclusion

We have examined the conceptual models behind the estimation of the intraclass correlation in particular, and of variance components in general. The models make underlying assumptions about the data to which they are applied. We have discussed how some of those assumptions may not hold under specific circumstances, thereby forcing us to qualify any inferences made based on our findings. We would therefore want to investigate two major aspects related to the proper application of the ANOVA models to data such as PIRLS:

- How does the unbalanced nature of PIRLS data affect the proper estimation of variance components in an ANOVA model?

- Can we be assured that the assumption of equal variance for the between-classroom within school variance component holds? If not, how does it affect the proper estimation of this variance component?

Based on our investigation of the intraclass correlations from the PIRLS data, we would be wise to consider intraclass correlations using classrooms as the first-level units, rather than schools, if they are to be used to compare countries. The resulting intraclass correlation is more readily comparable among participating countries. This is not to say that intraclass correlations based on schools as first-level units are inappropriate in general. Those can still be used for sample size determination, for example.

The application of a four-level ANOVA model to the PIRLS data has provided very useful insight into the national stratification strategies employed by the various participating countries. We are in a position to determine where successful strategies have been implemented and, more importantly, where alternate strategies would prove beneficial in future survey cycles. This latter point is particularly crucial if we consider that countries with large intraclass correlations tend to have low levels of achievement. Having a better understanding of the variance structure in those countries would be highly relevant both for policy implications and sample design development.

REFERENCES

BORTZ, Juergen, "Statistik Fuer Sozialwissenschaftler", Springer-Verlag, 1989

COCHRAN, William G., "Sampling Techniques, Third Edition", John Wiley and Sons, 1977

MARTIN, Mick, MULLIS, Ina V.S., et al., "PIRLS 2001 International Report", Boston College, 2003

GONZALEZ, Eugenio, Kennedy, Ann, "PIRLS 2001 User Guide for the International Database", Boston College, 2003

NETER, John, & WASSERMAN, William, "Applied Linear Statistical models", Irwin, 1974