5

Estimation of Sampling Variability, Design Effects, and Effective Sample Sizes

Eugenio J. Gonzalez
Boston College

Pierre Foy
Statistics Canada

5.1 OVERVIEW

In order to derive parameter estimates of the distribution of student achievement in each country that were both accurate and cost-effective, TIMSS made use of probability sampling techniques to sample students from national student populations.[1] The statistics computed from these national probability samples were used to estimate population parameters. Because there is some uncertainty involved in generalizing from samples to populations, the important statistics in the TIMSS international reports (Beaton, A.E. et al., 1996; Beaton, A.E. et al., 1996; Martin, M.O. et al., 1997; Mullis, I.V.S. et al., 1997) are presented together with their standard errors, which are a measure of this uncertainty.

The TIMSS sampling design applies stratified multistage cluster-sampling techniques to the problem of selecting efficient and accurate samples of students while working with schools and classes. Such complex designs capitalize on the structure of the student population (i.e., students grouped in classes within schools) to derive student samples that permit efficient and economical data collection. However, complex sampling designs make the task of computing standard errors to quantify sampling variability more difficult.

When, as in TIMSS, the sampling design involves multistage cluster sampling, there are several options for the estimation of sampling error that avoid the assumption of simple random sampling (see Wolter, 1985). The jackknife repeated replication technique (JRR) was chosen for estimating sampling errors in TIMSS because it is computationally straightforward, and provides approximately unbiased estimates of the sampling errors of means, totals, and percentages in complex sample designs.

The particular variation on the JRR technique used in TIMSS is described in Johnson and Rust (1992). This method assumes that the primary sampling units (PSUs) can be paired in a manner consistent with the sample design, and each pair regarded as members of a pseudo-stratum for variance estimation purposes. Note that when using the JRR technique for the estimation of sampling variability, the approach will appropriately reflect the combined effect of the between- and within-PSU contributions to the sampling variance. The general use of the JRR entails systematically assigning pairs of schools to sampling zones, and the random selection of one of these schools to have its contribution doubled, and the other zeroed, so as to construct a number of "pseudo-replicates" of the original sample. The statistic of interest is computed once for all of

[1] See Foy, Rust, and Schleicher (1996) for details of the TIMSS sampling design.

the original sample, and once more for each of the pseudo-replicate samples. The variation between the estimates from each of the replicate samples and the original sample estimate is the jackknife estimate of the sampling error of the statistic. Specific applications of the jackknife method are also discussed in the chapters describing the reporting of student achievement in subject-matter content areas (Chapter 9) and the Test-Curriculum Matching Analysis (Chapter 10).

Although the jackknife was the standard method of computing sampling errors in TIMSS, where standard errors were required for medians the balanced repeated replication (BRR) method was used instead. BRR was chosen over the JRR method in this instance because it produces asymptotically more consistent estimates for order statistics such as medians and percentiles.

5.2 CONSTRUCTION OF SAMPLING ZONES FOR SAMPLING VARIANCE ESTIMATION

An important step in applying the JRR and the BRR techniques to the estimation of sampling variability consists of assigning the schools to implicit strata, also known as sampling zones. Since the sample design called for 150 schools, a maximum of 75 zones was expected within each country, with two schools per zone. These zones were constructed by sequentially pairing the sampled schools. Because schools were generally sorted by a set of implicit stratification variables, the resulting assignment to sampling zones takes advantage of any benefit due to this implicit stratification. In countries where more than 150 schools were sampled, it was sometimes necessary to combine two schools for variance estimation purposes before assigning them to a sampling zone.

Zones were constructed within design domains, or explicit strata. In cases where there was an odd number of schools in an explicit stratum, either by design or because of school-level nonresponse, the students in the remaining school were randomly divided to make up two "quasi" schools for the purposes of calculating the jackknife standard error. Each zone then consisted of a pair of schools or "quasi" schools. Table 5.1 shows the number of sampling zones by grade in each country.

5.3 COMPUTING SAMPLING VARIANCE USING THE JRR METHOD

The JRR algorithm used in TIMSS assumes that there are H sampling zones within each country, each one containing two sampled schools selected independently. When computing a statistic "t" from the sample for a country, the formula for the JRR variance estimate of the statistic t is then given by the following equation:

$$Var_{jrr}(t) = \sum_{h=1}^{H} [t(J_h) - t(S)]^2$$

where H is the number of pairs in the sample for the country. The term $t(S)$ corresponds to the statistic computed for the whole sample (computed with any specific weights that may have been used to compensate for the unequal probability of selection of the different elements in the sample or any other post-stratification weight). The element $t(J_h)$ denotes the same statistic using the hth jackknife replicate, computed for all cases

Table 5.1 Sampling Zones by Grade Level*

Country	Third Grade	Fourth Grade	Seventh Grade	Eighth Grade
Australia	74	74	74	74
Austria	68	68	65	66
Belgium (Fl)	-	-	71	71
Belgium (Fr)	-	-	60	60
Bulgaria	-	-	52	58
Canada	75	75	75	75
Colombia	-	-	71	71
Cyprus	74	74	55	55
Czech Republic	73	73	75	75
Denmark	-	-	75	75
England	67	67	64	64
France	-	-	67	68
Germany	-	-	69	69
Greece	75	75	75	75
Hong Kong	62	62	43	43
Hungary	75	75	75	75
Iceland	75	75	75	75
Iran, Islamic Rep.	75	75	75	75
Ireland	73	73	66	66
Israel	-	44	-	23
Japan	74	74	75	75
Korea	75	75	75	75
Kuwait	-	75	-	36
Latvia (LSS)	59	59	64	64
Lithuania	-	-	73	73
Netherlands	52	52	48	48
New Zealand	75	75	75	75
Norway	70	70	72	74
Portugal	72	72	71	71
Romania	-	-	72	72
Russian Federation	-	-	41	41
Scotland	65	65	64	64
Singapore	75	75	69	69
Slovak Republic	-	-	73	73
Slovenia	61	61	61	61
South Africa	-	-	66	66
Spain	-	-	75	75
Sweden	-	-	75	60
Switzerland	-	-	75	75
Thailand	75	75	74	74
United States	59	59	55	55

A dash (-) means the country did not participate at this grade level
* Third, fourth, seventh, and eighth grades in most countries.

except those in the *h*th stratum of the sample, removing all cases associated with one of the randomly selected units of the pair within the *h*th stratum, and including, twice, the elements associated with the other unit in the *h*th stratum. In practice, this is effectively accomplished by recoding to zero the weights for the cases of the element of the pair to be excluded from the replication, and multiplying by two the weights of the remaining element within the *h*th pair.

The computation of the JRR variance estimate for any statistic from the TIMSS database requires the computation of any statistic up to 76 times for any given country: once to obtain the statistic for the full sample, and up to 75 times to obtain the statistics for each of the jackknife replicates (J_h). The number of times a statistic needs to be computed for a given country depends on the number of implicit strata or sampling zones defined for that country.

Doubling and zeroing the weights of the selected units within the sampling zones is accomplished effectively with the creation of replicate weights which are then used in the calculations. Gonzalez and Smith (1997) provide examples of how this approach allows standard statistical software such as SAS or SPSS to be used to compute JRR estimates of sampling variability in TIMSS. The replicate weight approach requires the user to temporarily create a new set of weights for each pseudo-replicate sample. Each replicate weight is equal to *k* times the overall sampling weight, where *k* can take values of zero, one or two depending on whether or not the case is to be removed from the computation, left as it is, or have its weight doubled. The value of *k* for an individual student record for a given replicate depends on the assignment of the record to the specific PSU and zone.

Within each zone the members of the pair of schools are assigned an indicator (u_i), coded randomly to 1 or 0 so that one of the members of each pair had values of 1 on the variable u_i, and the remaining member a value of 0. This indicator determines whether the weights for the elements in the school in this zone are to be doubled or zeroed. The replicate weight ($W_h^{g,i,j}$) for the elements in a school assigned to zone *h* is computed as the product of k_h times their overall sampling weight, where k_h can take values of zero, one, or two depending on whether the school is to be omitted, be included with its usual weight, or have its weight doubled for the computation of the statistic of interest. In TIMSS, the replicate weights are not permanent variables, but are created temporarily by the sampling variance estimation program as a useful computing device.

When creating the replicate weights the following procedure was followed:

Each sampled student was assigned a vector of 75 weights or $W_h^{g,i,j}$, where *h* takes values from 1 to 75.

The value of $W_0^{g,i,j}$ is the overall sampling weight which is simply the product of the final school weight, the appropriate final classroom weight, and the appropriate final student weight as described in chapter 4.

The replicate weights for a single case were then computed as:

$$W_h^{g,i,j} = W_0^{g,i,j} * k_{hi} ,$$

where the variable k_h for an individual *i* takes the value $k_{hi} = 2*u_i$ if the record belongs to zone *h*, and $k_{hi} = 1$ otherwise.

In TIMSS, a total of 75 replicate weights were computed for each country regardless of the number of actual zones within the country. If a country had fewer than 75 zones, then the replicate weights W_h, where *h* was greater than the number of zones within the country, were each the same as the overall sampling weight. Although this involved some redundant computation, having 75 replicate weights for each country has no effect on the size of the error variance computed using the jackknife formula, but facilitated the computation of standard errors for a number of countries at one time.

Figure 5.1 shows example SAS and SPSS computer code used to compute standard errors in TIMSS. Further examples are given in Gonzalez and Smith (1997). Although standard errors presented in the international reports were computed using SAS programs developed at the International Study Center, they were also verified against results produced by the WesVarPC software (Westat, 1997). Results were compared with each other for accuracy.[2]

Figure 5.1 Computer Code in SAS and SPSS to Generate JRR Replicate Weights

```
SAS Computer Code
data a;
  set datafile ;
array rwt  rwt1 - rwt75 ;         * Replicate Weights  ;
do i=1 to 75;
     if  jkzone  <>i                then rwt(i) = weight * 1;
     if (jkzone  = i & jkindic = 1) then rwt(i) = weight * 2;
     if (jkzone  = i & jkindic = 0) then rwt(i) = weight * 0;
end;

SPSS Computer Code
vector rwgt(75).
loop #i = 1 to 75.
if (jkzone = #i and jkindic = 0)  rwgt(#i) = weight * 0.
if (jkzone = #i and jkindic = 1)  rwgt(#i) = weight * 2.
if (jkzone <>#i               )  rwgt(#i) = weight * 1.
end loop.
```

5.4 COMPUTING SAMPLING VARIANCE USING THE BRR METHOD

Like the JRR method, balanced repeated replication (BRR) uses the variation between PSUs to estimate the sampling variation of a statistic. BRR forms a series of replicate half-samples by randomly selecting one of the pair of PSUs in each sampling zone. The weights of the selected PSUs are doubled to compensate for the omitted PSUs. When a statistic is computed independently from each of the replicate half-samples, the variation in the results may be used to estimate the sampling variance of that statistic. When computing a statistic *t* from the sample, the formula for the BRR variance estimate of the statistic *t* is given by the equation:

$$Var_{brr}(t) = \frac{\sum_{g=1}^{G} \mathrm{t}[t(B_g) - t(S)]^2}{G}$$

[2] Minor differences were occasionally found between the results obtained with WesVar and those obtained with software developed in-house. However, these differences were in all cases due to the fact that the two programs did not always choose the same PSUs in forming jackknife replicates. When identical jackknife replicates were used for both programs, the results were identical.

where G is the number of replicate half-samples formed from the entire sample. The term $t(S)$ corresponds to the statistic computed for the whole sample weighted to compensate for unequal selection probabilities and post-stratification adjustments. The element $t(B_g)$ denotes the same statistic using the gth replicate half-sample, formed by including only half the units in the original sample.

Although each replicate half-sample contains only one unit from each of the H strata, there are $2H$ possible half-samples for a given sample. When the number of strata, H, is large, the number of possible half-samples becomes enormous (3.78×10^{22} in the case of TIMSS with 75 replicates), and the computation of estimates of sampling variability using all such half-samples is no longer feasible. However, by selecting a subsample of G orthogonally balanced half-replicates it is possible to obtain an unbiased estimate of the variance that would have been obtained if all possible replicate half-samples had been used (see Wolter, 1985). This is true whenever G is an integral multiple of 4 that is greater than H, where H is the number of strata in the sample. The selection of the G half-samples is facilitated by the use of Hadamard matrices. For the purpose of computing the standard errors of medians for selected age groups in TIMSS, a Hadamard matrix of order 76 was used. The WesVarPC (Westat, 1997) software was used to construct the replicate half-samples in TIMSS, although the BRR sampling errors themselves were computed using software developed at the TIMSS International Study Center.

5.5 DESIGN EFFECTS AND EFFECTIVE SAMPLE SIZES

Complex survey samples such as those in TIMSS typically have sampling errors much larger than a simple random sample of the same size. This is because the elements of the clusters that are the building blocks of complex samples (in TIMSS the elements are students grouped in classes within schools) usually resemble each other more than they do members of the population in general. Consequently, a sample of size n drawn using simple random sampling from a population will usually be more efficient (i.e., have smaller sampling errors) than a sample of the same size drawn by means of a sample of pre-existing clusters in the population. The degree to which members of a cluster resemble each other more than they do elements of the population in general on some criterion variable may be measured by the intra-class correlation coefficient (Kish, 1965). When the intra-class correlation for a variable in a population is large, it may be necessary to select a much larger sample using cluster-sample techniques than would be necessary using simple random sampling methods.

Although the design efficiency of a multistage cluster sample is generally less than that of a simple random sample of the same size, multistage samples have other advantages in terms of economy and operational efficiency that make them the method of choice for surveys of student populations such as TIMSS. One way to quantify the reduction in design efficiency is through the design effect (Kish, 1965). The design effect for a variable is the ratio of two estimates of the sampling variance for a particular sample statistic: one computed using a technique such as the jackknife that takes all components of variance in the sampling design into account, and the other computed using the simple random sampling formula. The design effect is specific to the statistic and the variable for which it is computed. Since in TIMSS the technique for estimating sampling variance for means and percentages was the JRR, the design effect for these statistics was computed as the ratio of the JRR variance estimate to the variance estimate computed under the assumptions of simple random sampling. The design effect was computed as follows:

$$DEff(t) = \frac{Var_{jrr}(t)}{Var_{srs}(t)}$$

where $Var_{jrr}(t)$ is the sampling variance computed using the JRR method, and $Var_{srs}(t)$ is the variance computed under the assumptions of simple random sampling. When computing the design effect for the proportion of students (p) responding correctly to an item,[3] the sampling variance of the statistic ($Var_{srs}(P)$) based on a sample with n cases, was computed as:

$$Var_{srs}(P) = \frac{P * (1 - P)}{n}$$

When computing the design effect of a mean ($\bar{x}$), the sampling variance of the statistic ($Var_{srs}(\bar{x})$) based on a simple random sample with n cases was computed as:

$$Var_{srs}(\bar{x}) = \frac{Var_x}{n}$$

Another, related, measure of the design efficiency is the effective sample size. The effective sample size is the ratio of the actual sample size to the design effect. It is the number of sampling elements that would be required in a simple random sample to provide the same precision obtained with the actual complex sampling design. The effective sample size is computed as:

$$EffN(t) = \frac{N}{DEff(t)}$$

The TIMSS standard for sampling precision required that all student samples have an effective sample size of at least 400 for the main criterion variables (Foy, Rust, and Schleicher, 1996). Note that these requirements were for the entire populations (i.e., grades three and four combined for Population 1, and grades seven and eight for Population 2). Design effects and effective sample sizes for the mean mathematics and science achievement scores by population are presented in Tables 5.2 through 5.13. Design effects and effective sample sizes by grade and by grade and gender are included in Appendix C.

[3] Proportion correct is defined here as the proportion of students obtaining the maximum score on the item.

Table 5.2 Design Effects and Effective Sample Sizes for Third and Fourth Grades* (Combined) - Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	11248	516	9247.0	3.4	0.9	14.33	785
Austria	5171	524	7837.9	3.6	1.2	8.74	591
Canada	16002	502	7548.0	2.5	0.7	12.99	1232
Cyprus	6684	467	8028.1	2.5	1.1	5.20	1285
Czech Republic	6524	533	8376.5	2.8	1.1	6.10	1069
England	6182	485	8766.2	2.5	1.2	4.28	1445
Greece	6008	461	8703.9	3.4	1.2	8.02	749
Hong Kong	8807	556	6743.9	3.3	0.9	14.29	616
Hungary	6044	512	9176.7	3.4	1.2	7.63	792
Iceland	3507	442	5888.7	2.6	1.3	4.11	854
Iran, Islamic Rep.	6746	404	5179.4	3.4	0.9	15.44	437
Ireland	5762	513	8301.7	3.2	1.2	7.31	789
Israel	2351	531	7151.4	3.5	1.7	4.13	569
Japan	8612	568	7006.7	1.6	0.9	3.08	2795
Korea	5589	586	5812.0	1.9	1.0	3.32	1682
Kuwait	4318	400	4458.9	2.8	1.0	7.42	582
Latvia (LSS)	4270	498	7860.5	3.9	1.4	8.19	521
Netherlands	5314	535	6348.6	2.9	1.1	7.12	746
New Zealand	4925	470	8295.9	4.0	1.3	9.29	530
Norway	4476	462	6931.8	2.6	1.2	4.44	1009
Portugal	5503	452	7466.2	3.1	1.2	7.13	772
Scotland	6433	489	8128.2	3.2	1.1	8.20	784
Singapore	14169	588	11743.3	4.1	0.9	20.47	692
Slovenia	5087	520	7439.5	2.8	1.2	5.41	941
Thailand	5862	467	5482.5	4.4	1.0	20.46	287
United States	11115	512	8022.6	2.8	0.8	11.00	1010

*Third and fourth grades in most countries.

Table 5.3 Design Effects and Effective Sample Sizes for Third Grade* Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	4741	484	8114.9	4.0	1.3	9.55	497
Austria	2526	487	6877.0	5.3	1.6	10.50	241
Canada	7594	469	6111.8	2.7	0.9	8.75	868
Cyprus	3308	430	5984.4	2.8	1.3	4.23	782
Czech Republic	3256	497	6853.4	3.3	1.5	5.23	622
England	3056	456	7634.3	3.0	1.6	3.67	833
Greece	2955	428	7254.6	4.0	1.6	6.36	464
Hong Kong	4396	524	5250.2	3.0	1.1	7.74	568
Hungary	3038	476	7980.5	4.2	1.6	6.78	448
Iceland	1698	410	4519.7	2.8	1.6	2.93	579
Iran, Islamic Rep.	3361	378	4302.7	3.5	1.1	9.77	344
Ireland	2889	476	6558.0	3.6	1.5	5.71	506
Japan	4306	538	5671.4	1.5	1.1	1.76	2452
Korea	2777	561	4922.8	2.3	1.3	2.95	940
Latvia (LSS)	2054	463	6544.7	4.3	1.8	5.72	359
Netherlands	2790	493	4209.3	2.7	1.2	4.90	569
New Zealand	2504	440	6771.7	4.0	1.6	6.01	417
Norway	2219	421	5116.7	3.1	1.5	4.11	540
Portugal	2650	425	7293.0	3.8	1.7	5.24	506
Scotland	3132	458	6321.9	3.4	1.4	5.60	559
Singapore	7030	552	9984.8	4.8	1.2	16.22	433
Slovenia	2521	488	5980.9	2.9	1.5	3.59	701
Thailand	2870	444	5075.9	5.1	1.3	14.61	196
United States	3819	480	6709.8	3.4	1.3	6.56	582

*Third grade in most countries.

Table 5.4 Design Effects and Effective Sample Sizes for Fourth Grade* Mathematics Mean Scale Score - Population 1

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	6507	547	8399.9	3.2	1.1	7.93	820
Austria	2645	559	6212.5	3.1	1.5	4.05	653
Canada	8408	532	7000.5	3.3	0.9	13.11	641
Cyprus	3376	502	7461.4	3.1	1.5	4.43	761
Czech Republic	3268	567	7446.4	3.3	1.5	4.68	698
England	3126	513	8316.7	3.2	1.6	3.91	800
Greece	3053	492	8088.6	4.4	1.6	7.18	425
Hong Kong	4411	587	6240.4	4.3	1.2	13.11	336
Hungary	3006	548	7762.9	3.7	1.6	5.38	559
Iceland	1809	474	5232.1	2.7	1.7	2.50	725
Iran, Islamic Rep.	3385	429	4773.5	4.0	1.2	11.15	304
Ireland	2873	550	7283.4	3.4	1.6	4.68	614
Israel	2351	531	7151.4	3.5	1.7	4.13	569
Japan	4306	597	6590.6	2.1	1.2	2.80	1540
Korea	2812	611	5457.7	2.1	1.4	2.31	1219
Kuwait	4318	400	4458.9	2.8	1.0	7.42	582
Latvia (LSS)	2216	525	7199.9	4.8	1.8	7.15	310
Netherlands	2524	577	4974.4	3.4	1.4	5.74	440
New Zealand	2421	499	8022.9	4.3	1.8	5.60	432
Norway	2257	502	5497.9	3.0	1.6	3.61	624
Portugal	2853	475	6450.9	3.5	1.5	5.49	520
Scotland	3301	520	7994.1	3.9	1.6	6.25	528
Singapore	7139	625	10854.0	5.3	1.2	18.54	385
Slovenia	2566	552	6797.1	3.2	1.6	3.84	669
Thailand	2992	490	4834.7	4.7	1.3	13.59	220
United States	7296	545	7243.8	3.0	1.0	9.23	790

*Fourth grade in most countries.

Table 5.5 Design Effects and Effective Sample Sizes for Third and Fourth Grades* (Combined) Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	11248	537	9809.8	3.3	0.9	12.33	913
Austria	5171	536	7904.7	3.4	1.2	7.35	704
Canada	16002	521	8434.2	2.2	0.7	9.41	1700
Cyprus	6684	445	6461.3	2.4	1.0	6.07	1101
Czech Republic	6524	526	7859.0	2.8	1.1	6.36	1025
England	6182	525	10343.8	2.5	1.3	3.75	1647
Greece	6008	472	7503.3	3.3	1.1	8.75	687
Hong Kong	8807	508	6399.1	3.0	0.9	12.06	730
Hungary	6044	498	8322.2	3.3	1.2	7.94	761
Iceland	3507	470	8176.1	3.0	1.5	3.86	908
Iran, Islamic Rep.	6746	387	6567.5	3.6	1.0	13.42	503
Ireland	5762	510	8360.8	3.3	1.2	7.53	765
Israel	2351	505	7450.2	3.6	1.8	4.19	561
Japan	8612	548	5956.0	1.4	0.8	2.64	3263
Korea	5589	575	5353.3	1.7	1.0	3.16	1767
Kuwait	4318	401	7250.5	3.1	1.3	5.86	737
Latvia (LSS)	4270	491	7474.7	4.1	1.3	9.47	451
Netherlands	5314	528	5008.0	2.8	1.0	8.12	654
New Zealand	4925	503	10495.7	4.8	1.5	10.65	463
Norway	4476	491	9347.5	2.8	1.4	3.82	1171
Portugal	5503	453	8861.4	3.5	1.3	7.43	740
Scotland	6433	510	9546.3	3.8	1.2	9.59	671
Singapore	14169	517	10473.8	4.1	0.9	23.01	616
Slovenia	5087	516	6797.7	2.8	1.2	5.71	891
Thailand	5862	452	5923.1	5.2	1.0	27.15	216
United States	11115	538	9646.5	2.8	0.9	9.34	1190

*Third and fourth grades in most countries.

Table 5.6 Design Effects and Effective Sample Sizes for Third Grade* Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	4741	510	9561.3	4.4	1.4	9.54	497
Austria	2526	505	7667.5	4.6	1.7	7.06	358
Canada	7594	490	7766.0	2.5	1.0	6.31	1203
Cyprus	3308	415	5344.5	2.5	1.3	3.91	846
Czech Republic	3256	494	7156.4	3.4	1.5	5.35	609
England	3056	499	10118.3	3.5	1.8	3.63	842
Greece	2955	446	6800.1	3.9	1.5	6.70	441
Hong Kong	4396	482	5408.7	3.3	1.1	8.72	504
Hungary	3038	464	7886.0	4.1	1.6	6.35	478
Iceland	1698	435	6738.7	3.3	2.0	2.70	630
Iran, Islamic Rep.	3361	356	5772.2	4.2	1.3	10.14	331
Ireland	2889	479	7703.0	3.7	1.6	5.03	574
Japan	4306	522	5272.6	1.6	1.1	2.00	2156
Korea	2777	553	5103.3	2.4	1.4	3.14	885
Latvia (LSS)	2054	465	6817.4	4.5	1.8	6.20	331
Netherlands	2790	499	4022.8	3.2	1.2	7.01	398
New Zealand	2504	473	9913.8	5.2	2.0	6.87	365
Norway	2219	450	8069.1	3.9	1.9	4.12	538
Portugal	2650	423	9146.9	4.3	1.9	5.35	496
Scotland	3132	484	9021.1	4.2	1.7	6.19	506
Singapore	7030	488	9762.8	5.0	1.2	18.34	383
Slovenia	2521	487	6091.0	2.8	1.6	3.23	780
Thailand	2870	433	6010.7	6.6	1.4	20.63	139
United States	3819	511	8796.1	3.2	1.5	4.42	863

*Third grade in most countries.

Table 5.7 Design Effects and Effective Sample Sizes for Fourth Grade* Science Mean Scale Score - Population 1

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	6507	563	8699.4	3.0	1.2	6.78	960
Austria	2645	565	6370.7	3.3	1.6	4.43	597
Canada	8408	549	7381.8	3.0	0.9	10.14	829
Cyprus	3376	475	5730.1	3.3	1.3	6.44	524
Czech Republic	3268	557	6598.4	3.1	1.4	4.77	685
England	3126	551	9207.8	3.3	1.7	3.65	857
Greece	3053	497	6888.4	4.1	1.5	7.30	418
Hong Kong	4411	533	6046.9	3.7	1.2	10.03	440
Hungary	3006	532	6505.4	3.4	1.5	5.47	550
Iceland	1809	505	7207.9	3.3	2.0	2.74	660
Iran, Islamic Rep.	3385	416	5546.6	3.9	1.3	9.40	360
Ireland	2873	539	7205.7	3.3	1.6	4.41	651
Israel	2351	505	7450.2	3.6	1.8	4.19	561
Japan	4306	574	5296.3	1.8	1.1	2.53	1703
Korea	2812	597	4639.3	1.9	1.3	2.10	1342
Kuwait	4318	401	7250.5	3.1	1.3	5.86	737
Latvia (LSS)	2216	512	7022.1	4.9	1.8	7.65	290
Netherlands	2524	557	4319.8	3.1	1.3	5.45	463
New Zealand	2421	531	9418.7	4.9	2.0	6.14	394
Norway	2257	530	7432.4	3.6	1.8	3.85	586
Portugal	2853	480	7122.1	4.0	1.6	6.46	441
Scotland	3301	536	8731.0	4.2	1.6	6.58	501
Singapore	7139	547	9445.0	5.0	1.2	19.12	373
Slovenia	2566	546	5780.5	3.3	1.5	4.96	517
Thailand	2992	473	5012.2	4.9	1.3	14.26	210
United States	7296	565	9028.6	3.1	1.1	7.65	954

*Fourth grade in most countries.

Table 5.8 Design Effects and Effective Sample Sizes for Seventh and Eighth Grades* (Combined) Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	12,852	514	9,287.0	3.5	0.9	17.27	744
Austria	5,786	524	8,080.8	2.5	1.2	4.50	1,285
Belgium (Fl)	5,662	562	7,270.7	4.0	1.1	12.16	465
Belgium (Fr)	4,883	518	6,907.2	3.0	1.2	6.31	774
Bulgaria	3,771	527	11,612.4	4.6	1.8	6.97	541
Canada	16,581	511	7,196.6	1.9	0.7	8.42	1,970
Colombia	5,304	376	4,103.4	2.8	0.9	10.25	518
Cyprus	5,852	459	7,394.3	1.4	1.1	1.55	3,770
Czech Republic	6,672	544	8,778.7	3.8	1.1	11.00	606
Denmark	4,370	485	6,911.4	1.9	1.3	2.32	1,885
England	3,579	491	8,587.4	2.4	1.5	2.40	1,493
France	6,014	514	6,136.6	2.4	1.0	5.51	1,091
Germany	5,763	497	7,780.5	4.1	1.2	12.41	464
Greece	7,921	461	8,019.5	2.6	1.0	6.91	1,146
Hong Kong	6,752	576	10,163.8	6.8	1.2	30.29	223
Hungary	5,978	519	8,745.0	3.0	1.2	6.34	943
Iceland	3,730	473	5,376.0	2.6	1.2	4.60	811
Iran, Islamic Rep.	7,429	414	3,551.4	1.8	0.7	6.59	1,127
Ireland	6,203	513	8,239.7	3.4	1.2	8.59	722
Israel	1,415	522	8,463.5	6.2	2.4	6.36	222
Japan	10,271	588	10,102.3	1.7	1.0	2.88	3,567
Korea	5,827	592	11,622.5	2.0	1.4	2.06	2,827
Kuwait	1,655	392	3,325.4	2.5	1.4	3.15	526
Latvia (LSS)	4,976	477	6,531.0	2.4	1.1	4.55	1,095
Lithuania	5,056	454	6,656.9	2.8	1.1	5.82	869
Netherlands	4,084	529	7,257.6	4.6	1.3	12.14	336
New Zealand	6,867	490	8,180.3	2.9	1.1	7.28	943
Norway	5,736	482	6,855.2	1.9	1.1	3.16	1,818
Portugal	6,753	438	4,058.8	2.0	0.8	6.71	1,007
Romania	7,471	468	7,709.6	3.3	1.0	10.49	712
Russian Federation	8,160	518	8,399.0	3.9	1.0	14.71	555
Scotland	5,776	481	7,481.5	4.1	1.1	13.19	438
Singapore	8,285	622	8,682.6	4.8	1.0	22.21	373
Slovak Republic	7,101	527	8,230.6	2.7	1.1	6.37	1,115
Slovenia	5,606	519	7,642.8	2.4	1.2	4.40	1,274
South Africa	9,792	351	4,167.8	3.1	0.7	23.21	422
Spain	7,596	468	5,504.4	1.9	0.9	4.83	1,574
Sweden	6,906	498	7,024.7	2.0	1.0	3.82	1,808
Switzerland	8,940	526	7,097.2	2.1	0.9	5.39	1,658
Thailand	11,643	508	6,952.1	4.9	0.8	40.70	286
United States	10,973	488	8,261.9	4.3	0.9	24.83	442

*Seventh and eighth grades in most countries.

Table 5.9 Design Effects and Effective Sample Sizes for Seventh Grade* Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	5,599	498	8,437.6	3.8	1.2	9.59	584
Austria	3,013	509	7,260.4	3.0	1.6	3.70	815
Belgium (Fl)	2,768	558	5,877.2	3.5	1.5	5.91	469
Belgium (Fr)	2,292	507	6,085.4	3.5	1.6	4.73	484
Bulgaria	1,798	514	10,670.8	7.5	2.4	9.39	191
Canada	8,219	494	6,396.9	2.2	0.9	6.30	1,305
Colombia	2,655	369	3,967.1	2.7	1.2	4.89	543
Cyprus	2,929	446	6,747.6	1.9	1.5	1.61	1,823
Czech Republic	3,345	523	7,972.0	4.9	1.5	10.15	329
Denmark	2,073	465	6,030.0	2.1	1.7	1.56	1,330
England	1,803	476	8,084.6	3.7	2.1	2.98	606
France	3,016	492	5,460.0	3.1	1.3	5.46	552
Germany	2,893	484	7,237.0	4.1	1.6	6.77	428
Greece	3,931	440	7,289.8	2.8	1.4	4.34	905
Hong Kong	3,413	564	9,841.0	7.8	1.7	21.34	160
Hungary	3,066	502	8,232.0	3.7	1.6	5.01	613
Iceland	1,957	459	4,594.9	2.6	1.5	2.84	689
Iran, Islamic Rep.	3,735	401	3,232.4	2.0	0.9	4.59	815
Ireland	3,127	500	7,537.8	4.1	1.6	7.03	445
Japan	5,130	571	9,220.1	1.9	1.3	2.05	2,507
Korea	2,907	577	10,930.5	2.5	1.9	1.72	1,689
Latvia (LSS)	2,567	462	5,859.6	2.8	1.5	3.45	743
Lithuania	2,531	428	5,657.0	3.2	1.5	4.45	568
Netherlands	2,097	516	6,231.6	4.1	1.7	5.66	370
New Zealand	3,184	472	7,540.2	3.8	1.5	6.08	523
Norway	2,469	461	5,779.8	2.8	1.5	3.42	721
Portugal	3,362	423	3,569.6	2.2	1.0	4.62	727
Romania	3,746	454	7,091.3	3.4	1.4	5.99	625
Russian Federation	4,138	501	7,781.8	4.0	1.4	8.30	499
Scotland	2,913	463	6,670.6	3.7	1.5	6.06	480
Singapore	3,641	601	8,694.2	6.3	1.5	16.88	216
Slovak Republic	3,600	508	7,240.7	3.4	1.4	5.66	636
Slovenia	2,898	498	6,715.2	3.0	1.5	3.77	769
South Africa	5,301	348	4,023.3	3.8	0.9	19.06	278
Spain	3,741	448	4,836.5	2.2	1.1	3.87	968
Sweden	2,831	477	5,911.6	2.5	1.4	2.93	965
Switzerland	4,085	506	5,684.3	2.3	1.2	3.79	1,078
Thailand	5,810	495	6,178.2	4.9	1.0	22.14	262
United States	3,886	476	7,966.0	5.5	1.4	14.73	264

*Seventh grade in most countries.

Table 5.10 Design Effects and Effective Sample Sizes for Eighth Grade* Mathematics Mean Scale Score - Population 2

Country	Sample Size	Mean Mathematics Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	7,253	530	9,651.1	4.0	1.2	12.18	596
Austria	2,773	539	8,462.9	3.0	1.7	3.05	910
Belgium (Fl)	2,894	565	8,435.6	5.7	1.7	11.00	263
Belgium (Fr)	2,591	526	7,431.9	3.4	1.7	4.03	644
Bulgaria	1,973	540	12,187.6	6.3	2.5	6.42	308
Canada	8,362	527	7,444.2	2.4	0.9	6.51	1,285
Colombia	2,649	385	4,120.9	3.4	1.2	7.64	347
Cyprus	2,923	474	7,684.9	1.9	1.6	1.36	2,155
Czech Republic	3,327	564	8,771.2	4.9	1.6	9.21	361
Denmark	2,297	502	7,007.4	2.8	1.7	2.61	879
England	1,776	506	8,641.6	2.6	2.2	1.44	1,234
France	2,998	538	5,781.2	2.9	1.4	4.33	693
Germany	2,870	509	8,025.5	4.5	1.7	7.22	398
Greece	3,990	484	7,798.5	3.1	1.4	4.81	829
Hong Kong	3,339	588	10,188.4	6.5	1.7	13.94	239
Hungary	2,912	537	8,641.1	3.2	1.7	3.52	826
Iceland	1,773	487	5,780.1	4.5	1.8	6.31	281
Iran, Islamic Rep.	3,694	428	3,513.5	2.2	1.0	4.88	758
Ireland	3,076	527	8,564.1	5.1	1.7	9.47	325
Israel	1,415	522	8,463.5	6.2	2.4	6.36	222
Japan	5,141	605	10,388.5	1.9	1.4	1.74	2,951
Korea	2,920	607	11,848.0	2.4	2.0	1.40	2,091
Kuwait	1,655	392	3,325.4	2.5	1.4	3.15	526
Latvia (LSS)	2,409	493	6,743.4	3.1	1.7	3.50	688
Lithuania	2,525	477	6,424.9	3.5	1.6	4.91	515
Netherlands	1,987	541	7,897.7	6.7	2.0	11.15	178
New Zealand	3,683	508	8,153.3	4.5	1.5	9.08	406
Norway	3,267	503	7,033.6	2.2	1.5	2.20	1,487
Portugal	3,391	454	4,075.6	2.5	1.1	5.15	659
Romania	3,725	482	7,958.2	4.0	1.5	7.63	488
Russian Federation	4,022	535	8,446.6	5.3	1.4	13.48	298
Scotland	2,863	498	7,639.1	5.5	1.6	11.25	254
Singapore	4,644	643	7,782.4	4.9	1.3	14.39	323
Slovak Republic	3,501	547	8,474.6	3.3	1.6	4.51	776
Slovenia	2,708	541	7,700.1	3.1	1.7	3.36	806
South Africa	4,491	354	4,270.1	4.4	1.0	20.79	216
Spain	3,855	487	5,397.9	2.0	1.2	2.87	1,341
Sweden	4,075	519	7,278.7	3.0	1.3	4.90	832
Switzerland	4,855	545	7,670.4	2.8	1.3	4.88	996
Thailand	5,833	522	7,365.0	5.7	1.1	25.79	226
United States	7,087	500	8,266.4	4.6	1.1	18.45	384

*Eighth grade in most countries.

Table 5.11 Design Effects and Effective Sample Sizes for Seventh and Eighth Grades* (Combined) - Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	12,852	524	11,329.0	3.3	0.9	12.28	1,046
Austria	5,786	538	9,606.7	2.9	1.3	5.03	1,150
Belgium (Fl)	5,662	540	6,125.6	2.6	1.0	6.16	920
Belgium (Fr)	4,883	458	7,000.1	2.5	1.2	4.48	1,091
Bulgaria	3,771	548	11,746.9	4.0	1.8	5.22	722
Canada	16,581	515	8,596.0	2.0	0.7	7.40	2,239
Colombia	5,304	398	5,580.2	3.4	1.0	11.05	480
Cyprus	5,852	440	8,152.7	1.3	1.2	1.18	4,956
Czech Republic	6,672	553	7,549.6	2.7	1.1	6.68	999
Denmark	4,370	460	7,993.3	2.1	1.4	2.39	1,832
England	3,579	532	11,125.7	2.6	1.8	2.18	1,641
France	6,014	474	6,229.8	2.1	1.0	4.16	1,446
Germany	5,763	515	9,962.9	4.1	1.3	9.63	599
Greece	7,921	472	8,025.1	2.1	1.0	4.45	1,781
Hong Kong	6,752	509	7,870.6	4.6	1.1	18.14	372
Hungary	5,978	535	8,551.7	2.6	1.2	4.68	1,277
Iceland	3,730	478	6,195.1	2.5	1.3	3.89	959
Iran, Islamic Rep.	7,429	452	5,474.7	2.1	0.9	6.26	1,187
Ireland	6,203	516	9,161.1	3.0	1.2	6.03	1,028
Israel	1,415	524	10,758.9	5.7	2.8	4.33	327
Japan	10,271	552	8,175.0	1.6	0.9	3.13	3,285
Korea	5,827	550	8,821.1	1.7	1.2	1.97	2,958
Kuwait	1,655	430	5,459.9	3.7	1.8	4.18	396
Latvia (LSS)	4,976	459	6,945.4	2.1	1.2	3.13	1,591
Lithuania	5,056	441	7,788.4	2.8	1.2	5.14	983
Netherlands	4,084	540	7,216.3	3.6	1.3	7.43	550
New Zealand	6,867	504	10,140.0	3.0	1.2	5.97	1,150
Norway	5,736	505	7,894.2	1.8	1.2	2.26	2,539
Portugal	6,753	453	5,940.1	2.0	0.9	4.63	1,459
Romania	7,471	469	10,470.0	4.1	1.2	12.20	612
Russian Federation	8,160	510	9,710.2	3.6	1.1	10.92	747
Scotland	5,776	493	9,984.8	4.1	1.3	9.80	589
Singapore	8,285	576	10,542.6	5.3	1.1	21.76	381
Slovak Republic	7,101	527	8,127.0	2.7	1.1	6.14	1,157
Slovenia	5,606	544	7,762.2	2.0	1.2	2.78	2,019
South Africa	9,792	322	9,192.8	4.6	1.0	22.80	429
Spain	7,596	497	6,627.9	1.7	0.9	3.23	2,353
Sweden	6,906	512	8,184.2	2.0	1.1	3.45	2,000
Switzerland	8,940	503	7,867.9	1.9	0.9	4.30	2,078
Thailand	11,643	509	5,266.7	3.1	0.7	21.79	534
United States	10,973	521	11,268.9	4.6	1.0	20.22	543

*Seventh and eighth grades in most countries.

Table 5.12 Design Effects and Effective Sample Sizes for Seventh Grade* Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	5,599	504	10,522.1	3.6	1.4	6.78	826
Austria	3,013	519	8,833.5	3.1	1.7	3.36	897
Belgium (Fl)	2,768	529	5,343.3	2.6	1.4	3.37	821
Belgium (Fr)	2,292	442	6,183.9	3.0	1.6	3.45	665
Bulgaria	1,798	531	10,607.9	5.4	2.4	5.02	358
Canada	8,219	499	8,045.0	2.3	1.0	5.46	1,505
Colombia	2,655	387	5,218.9	3.2	1.4	5.34	497
Cyprus	2,929	420	7,567.9	1.8	1.6	1.31	2,238
Czech Republic	3,345	533	6,684.3	3.3	1.4	5.56	602
Denmark	2,073	439	7,453.4	2.1	1.9	1.28	1,625
England	1,803	512	10,226.4	3.5	2.4	2.16	834
France	3,016	451	5,510.5	2.6	1.4	3.62	833
Germany	2,893	499	9,147.1	4.1	1.8	5.19	557
Greece	3,931	449	7,631.1	2.6	1.4	3.38	1,163
Hong Kong	3,413	495	7,471.9	5.5	1.5	13.77	248
Hungary	3,066	518	8,351.8	3.2	1.7	3.69	830
Iceland	1,957	462	5,643.0	2.8	1.7	2.68	730
Iran, Islamic Rep.	3,735	436	5,124.9	2.6	1.2	4.77	784
Ireland	3,127	495	8,288.2	3.5	1.6	4.50	695
Japan	5,130	531	7,427.5	1.9	1.2	2.41	2,129
Korea	2,907	535	8,419.3	2.1	1.7	1.57	1,848
Latvia (LSS)	2,567	435	6,087.5	2.7	1.5	3.07	835
Lithuania	2,531	403	6,313.6	3.4	1.6	4.59	551
Netherlands	2,097	517	6,248.5	3.6	1.7	4.33	484
New Zealand	3,184	481	9,316.0	3.4	1.7	4.00	797
Norway	2,469	483	7,195.8	2.9	1.7	2.88	857
Portugal	3,362	428	5,109.1	2.1	1.2	2.91	1,155
Romania	3,746	452	9,999.2	4.4	1.6	7.30	513
Russian Federation	4,138	484	8,890.2	4.2	1.5	8.06	514
Scotland	2,913	468	8,773.3	3.8	1.7	4.85	601
Singapore	3,641	545	10,030.6	6.6	1.7	15.94	228
Slovak Republic	3,600	510	7,218.0	3.0	1.4	4.59	784
Slovenia	2,898	530	7,387.2	2.4	1.6	2.19	1,322
South Africa	5,301	317	8,470.9	5.3	1.3	17.46	304
Spain	3,741	477	6,387.0	2.1	1.3	2.65	1,410
Sweden	2,831	488	7,110.8	2.6	1.6	2.62	1,082
Switzerland	4,085	484	6,709.2	2.5	1.3	3.67	1,113
Thailand	5,810	493	4,779.5	3.0	0.9	10.85	536
United States	3,886	508	11,014.6	5.5	1.7	10.51	370

*Seventh grade in most countries.

Table 5.13 Design Effects and Effective Sample Sizes for Eighth Grade* Science Mean Scale Score - Population 2

Country	Sample Size	Mean Science Score	Variance	JRR s.e.	SRS s.e.	Design Effect	Effective Sample Size
Australia	7,253	545	11,338.8	3.9	1.3	9.50	763
Austria	2,773	558	9,636.0	3.7	1.9	3.87	717
Belgium (Fl)	2,894	550	6,579.3	4.2	1.5	7.62	380
Belgium (Fr)	2,591	471	7,315.2	2.8	1.7	2.87	904
Bulgaria	1,973	565	12,273.1	5.3	2.5	4.49	439
Canada	8,362	531	8,644.9	2.6	1.0	6.46	1,295
Colombia	2,649	411	5,703.8	4.1	1.5	7.68	345
Cyprus	2,923	463	7,838.6	1.9	1.6	1.38	2,112
Czech Republic	3,327	574	7,574.0	4.3	1.5	8.11	410
Denmark	2,297	478	7,741.4	3.1	1.8	2.91	790
England	1,776	552	11,202.9	3.3	2.5	1.78	999
France	2,998	498	5,893.4	2.5	1.4	3.15	952
Germany	2,870	531	10,284.8	4.8	1.9	6.45	445
Greece	3,990	497	7,220.9	2.2	1.3	2.75	1,448
Hong Kong	3,339	522	7,908.8	4.7	1.5	9.26	361
Hungary	2,912	554	8,105.2	2.8	1.7	2.81	1,036
Iceland	1,773	494	6,246.6	4.0	1.9	4.64	382
Iran, Islamic Rep.	3,694	470	5,277.5	2.4	1.2	4.02	919
Ireland	3,076	538	9,132.9	4.5	1.7	6.89	447
Israel	1,415	524	10,758.9	5.7	2.8	4.33	327
Japan	5,141	571	8,108.4	1.6	1.3	1.72	2,992
Korea	2,920	565	8,774.9	1.9	1.7	1.22	2,395
Kuwait	1,655	430	5,459.9	3.7	1.8	4.18	396
Latvia (LSS)	2,409	485	6,589.1	2.7	1.7	2.69	897
Lithuania	2,525	476	6,564.2	3.4	1.6	4.51	560
Netherlands	1,987	560	7,225.6	5.0	1.9	6.80	292
New Zealand	3,683	525	9,958.0	4.4	1.6	7.04	523
Norway	3,267	527	7,628.7	1.9	1.5	1.63	2,010
Portugal	3,391	480	5,447.4	2.3	1.3	3.41	993
Romania	3,725	486	10,345.6	4.7	1.7	8.10	460
Russian Federation	4,022	538	9,075.2	4.0	1.5	7.02	573
Scotland	2,863	517	9,968.9	5.1	1.9	7.48	383
Singapore	4,644	607	9,097.9	5.5	1.4	15.65	297
Slovak Republic	3,501	544	8,458.0	3.2	1.6	4.36	804
Slovenia	2,708	560	7,695.7	2.5	1.7	2.16	1,252
South Africa	4,491	326	9,769.0	6.6	1.5	20.29	221
Spain	3,855	517	6,072.4	1.7	1.3	1.84	2,096
Sweden	4,075	535	8,145.7	3.0	1.4	4.41	923
Switzerland	4,855	522	8,266.9	2.5	1.3	3.67	1,324
Thailand	5,833	525	5,232.6	3.7	0.9	15.67	372
United States	7,087	534	11,178.9	4.7	1.3	14.29	496

*Eighth grades in most countries.

REFERENCES

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1996a). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O. Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1996b). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Foy, P., Rust, K., and Schleicher, A. (1996). Sample Design. In M.O. Martin and D.L. Kelly (Eds.), *Third International Mathematics and Science Study technical report, volume I: Design and development*. Chestnut Hill, MA: Boston College.

Gonzalez, E. J. and Smith, T. A., Eds. (1997). *User guide for the TIMSS international database: Primary and middle school years – 1995 assessment.* Chestnut Hill, MA: Boston College.

Johnson, E. G. and Rust, K.F. (1992). Population references and variance estimation for NAEP data. *Journal of Educational Statistics*, 17, 175-190.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., and Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., and Smith, T.A. (1997). *Mathematics achievement in the primary school years IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Westat, Inc. (1997). *A user's guide to WesVarPC.* Rockville, MD: Westat, Inc.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.