/*

SAS Macro for Estimating Education-Study Statistical Power Parameters

George M. Harrison
Curriculum Research & Development Group
University of Hawaii at Manoa
Honolulu, HI

December 2012

Please cite as
Harrison, G. M. (2012). A SAS macro for estimating education-study statistical
power parameters [Computer software code]. Retrieved from http://goo.gl/moLFl

/**********DESCRIPTION**********

This code calculates the 95% ICC confidence interval of a two-level set of data that you plan to randomly sample from a larger known finite population of data. For example, if you have a dataset of your finite population, such as all the standardized test scores of students clustered within schools within a region, and you plan to randomly sample a set of clusters (schools) and Level 1 units (students) from that dataset in order to conduct your planned study, this code will help you to calculate the point estimate and upper bound of the unconditional (null model) ICC. The ICC confidence interval can then be used to estimate the 95% confidence interval of the minimal detectable effect size (MDES, also reported in the output) of your planned design.

The ICCs are estimated using PROC MIXED on the unconditional model with restricted maximum likelihood estimation. The formula for calculating the standard errors of the ICC estimates is from Fisher (1925) and presented in Jacob, Zhu, and Bloom (2010). The 95% confidence limits of the ICC estimates are calculated using the t-distribution (based on the degrees of freedom of the sampled number of schools) rather than the traditional 1.96 multiplier (which would assume a large sample size).

Additionally, this code allows you to calculate several estimates at the same time. That is, if your dataset has more than one outcome variable, such as Reading and Math test scores, and more than one category of data, such as grades of students, you can calculate these ICCs in a single run.

The MDES, using Jacob, Zhu, and Bloom's (2010) equation, given the point estimate and upper ICC are printed as well. If you know the cluster-level covariate r-square (such as school-level scores on a pretest), enter this in the design. For greater flexibility in estimating MDES (i.e., if you have more than one covariate), use Optimal Design software (Spybrook, Bloom, Congdon, Martinez, & Raudenbush, 2011).

The syntax has been tested in SAS 9.1.3 and 9.3. (It has not been tested in SAS Enterprise.)

Follow the instructions to specify your planned design.

/**********INSTRUCTIONS**********

1. Format the dataset containing your finite population data:
   a) Make sure your data are in a SAS dataset.
   b) If your cluster variable (such as school) is in character format, convert it to NUMERIC format
      [for example, make the first school "101", the second school "102" (without quotes), and so on].
   c) If you have different grade levels, and you wish to calculate separate ICC estimates for each grade, make sure this
      variable is in NUMERIC format and is one or two digits long (for example, third grade would be "3", fourth grade
      would be "4", and so on).
      IF YOU ARE NOT ESTIMATING ICCS FOR DIFFERENT GRADES (e.g., if your dataset is a single grade), add a dummy
      variable to your dataset for group (call it "grade" for example) and assign a single number (e.g., "1") to every
      observation.
2. Point the libname, ICCLib, to the directory containing your SAS dataset.
3. Specify the fields after the %LET statements. For example,
   a) The name of your population dataset (after the line beginning with "%LET PopDataSet =").
   b) The name of your cluster variable (e.g., "School") that is in your dataset.
   c) The name of your grade variable (or other category identifier used to conduct separate
      analyses, such as reduced-or-free-lunch status) in your dataset.
   d) The minimum number of students (or other Level 1 unit) within each cluster for that cluster to be included in the
      sampling. (For example, if your research plan is to randomly sample from among the schools that have at least 30
      students in the grade, specify "30" here.)
   e) The number of students per cluster you plan to include in your study (e.g., 25 students). Enter a period if you plan
      to include all possible students in the school, then make sure you enter "N" after the "%LET UseNumStudents = ".
   f) A code, if any, that you use for invalid scores (leave it as "." if you have no special code).
4. Specify your design by entering it into the dataset entitled Design. Make sure this dataset has a row for each possible
   combination in your design that you wish to examine. That is,
   a) the grade(one or two digits),
   b) outcome variable (maximum 8 characters long; distinguish outcome variables by their first 4 digits);
   c) number of clusters (schools) you plan to randomly sample from your population,
   d) the cluster-level (school-level) covariate r-square value (e.g., the r-square of the school-level covariate such as
      the previous year's test scores.)
      NOTE: ENTER "0" IF YOU ARE NOT INCLUDING A COVARIATE.
5. After you run the code, make sure all ICC estimates converged. For example, do a search in your log (use Ctrl + F) for
   "Did not converge" and "Stopped because of infinite likelihood".

```
/*********************************USER-DEFINED DATA SPECIFICATION*********************************/

LIBNAME ICCLib 'C:\. . .your data set directory'; /*<----Specify the location of your dataset.*/


%LET PopDataSet = Your_dataset_name; /*<--Enter the name of your dataset after the equals sign.*/
%LET ClusterVarbName= School;  /*<--Enter the name of your cluster variable (e.g., "School" or "Class").*/
%LET GradeVarbName  = grade;   /*<--Enter the name of your grade variable.*/
%LET MinNumStudnts  = 30;      /*<--Enter the minimum number of Level 1 observations (students) per cluster (school).*/
%LET NumStudents    = 25;      /*<--Enter the number of Level 1 observations (students) per cluster (school) you plan
                                  to randomly sample (e.g., if you plan to sample 25 students per school). NOTE. If you
                                  will use all students, enter "." and then make sure you specify "N" after
                                  %LET UseNumStudents= */
%LET UseNumStudents = Y;       /*<--Enter "Y" for YES, or "N" for NO.  If you wish to use the number of students per
                                  school in calculating the ICC Standard Error and the MDESs, enter "Y". If you wish
                                  to use the harmonic mean number of students in your datasets in your planned design,
                                  enter "N". (The harmonic mean is required if you will include all possible students
                                  in each school)*/
%LET PcntTreatGrp   = .5;      /*<--Percent of the cluster variable (schools) you plan to assign to the treatment group
                                  (e.g., enter .5 in designs where half of the schools are in the treatment, half are in
                                   the control condition).*/
%LET InvalidScore   = .;       /*<--If your data set has a code for invalid datapoints (e.g., if you use "9" or "100"
                                   to indicate missing), enter it here. If you do not use a code, enter a period */

/* ENTER YOUR PROPOSED DESIGN IN THE DATASET BELOW (Replace the existing example design dataset.)
   Note. If you do not have a covariate for a test, enter 0 (or .00) for the Cov variable. */
DATA Design;
INPUT  GradeLevel Test$ NumSchools Cov;
DATALINES;
3 Math     20    0
3 Math     30    .00
3 Math     30    .70
4 Reading  40    .64
6 Reading  20    .75
6 Reading  40    .00
6 Reading  40    .70
6 Reading  1000  .00
;
RUN;
/*********************************END OF USER-DEFINED DATA SPECIFICATION*********************************/
```

```
/*********************************************MACRO*************************************************************/

    /*Adds the t-distribution based on the number of schools and alpha at .05, two-tailed*/
ODS Graphics off;
DATA Design2;
   SET Design;
   tdist = TINV(1-(.05/2),(NumSchools-1));
   LENGTH Grade $2.;
   LENGTH Covlabel $2.;
   CovLabel = PUT((100*Cov),Z2.);
   Grade = PUT(GradeLevel,Z2.);
   IF LENGTH(Test)      = 1 THEN TestLabel = CATS(PUT(Test,$Char4.),"___");
    ELSE IF LENGTH(Test)= 2 THEN TestLabel = CATS(PUT(Test,$Char4.),"__");
    ELSE IF LENGTH(Test)= 3 THEN TestLabel = CATS(PUT(Test,$Char4.),"_");
    ELSE TestLabel = PUT(Test,$Char4.);
 RUN;


/*********************************MACRO CODE GENERATION*************************************/

    /*Creates the code (based on the user's specification above) that will later be used to
    invoke the macros.*/

PROC SQL NOPRINT;
    SELECT COUNT(*)
    INTO : nobs
    FROM Design2;
QUIT;
%LET nobs = %trim(%left(&nobs));

DATA Stringit;
    LENGTH BaseStrn$ 200;
    LENGTH CallSets$ 200;
    LENGTH CallICCs$ 200;
    LENGTH CallMerg$ 500;
   SET design2;
    InvokSe= %NRSTR('%Sets');
    InvokIc= %NRSTR('%ICC');
    InvokMe= %NRSTR('%MeanSamps');
    TestAbrev = SUBSTR(Test,1,4);
   DO i = 1 TO &nobs. BY 1;
   IF _N_ = i THEN
    BaseStrn =
Compress("GRADE"||"="||Grade||","||"TEST"||"="||Test||","||"TestAbrev"||"="||TestLabel||","||"SampSize"||"="||NumSchools||')');
    CallSets =  Compress(InvokSe||'('||BaseStrn);
    CallICCs =  Compress(InvokIc||'('||"tdist"||"="||tdist||","||"Cov="||Cov||","||"CovLabel="||CovLabel||","||BaseStrn);
    CallMerg =  Compress(TestLabel||Grade||NumSchools||CovLabel);
   END;
   DROP i InvokSe InvokIc InvokMe;
RUN;
```

```
PROC SQL NOPRINT;
    SELECT DISTINCT (CallSets)
    INTO : InvokeSets
    SEPARATED BY " "
    FROM work.Stringit (WHERE=(CallSets NE ' '));
    SELECT DISTINCT(CallICCs)
    INTO : InvokeICCs
    SEPARATED BY " "
    FROM work.Stringit (WHERE=(CallICCs NE ' '));
    SELECT DISTINCT(CallMerg)
    INTO : InvokeMerg
    SEPARATED BY " "
    FROM work.Stringit (WHERE=(CallMerg NE ' '));
QUIT;

    /*Renames L1 (student) and L2 (cluster) variables for use in subsequent macro*/
DATA AADataSet;
   LENGTH SCHL1D 8;
   LENGTH GRAD1D 8;
   SET ICCLib.&PopDataSet. ;
    SCHL1D = &ClusterVarbName. ;
    GRAD1D = &GradeVarbName.;
RUN;

    /*Specifies the number of Level 1 observations (students) to include in subsequent SE and MDES calculations */;
%MACRO Level1Specs;
    %GLOBAL Note;
    %GLOBAL NL1;
    %IF &UseNumStudents. = Y %THEN %DO;
       %LET NL1=&NumStudents.;
       %LET Note=You specified to calculate ICC SEs and MDESs based on &NumStudents. students per school.;
       %END;
      %ELSE %DO;
       %LET NL1=Harmean;
       %LET Note=You specified to calculate ICC SEs and MDESs based on the harmonic mean number of students per school.;
       %END;
%MEND Level1Specs;
%Level1Specs;
```

```
/**********************************SUBSET SELECTION*************************************

    Eliminates observations with no scores and of schools which have fewer than your specified number of scores per school
    (i.e., if you specified to include only schools with at least 25 students' scores.
    This part also creates subsets of DATA by grade.*/


%MACRO Sets ( GRADE= , Test=, TestAbrev= ,SampSize= );
DATA Raw&TestAbrev.Lev01_&GRADE. (KEEP = SCHL1D GRAD1D &test. );
   SET AADataSet;
   IF SCHL1D NE . and &test. NE . and &test. NE &InvalidScore. and GRAD1D = &GRADE. ;
RUN;
PROC SUMMARY DATA = Raw&TestAbrev.Lev01_&GRADE. NWAY;
   CLASS SCHL1D;
   OUTPUT  out = Raw&TestAbrev.Lev02temp_&GRADE.;
RUN;
DATA Raw&TestAbrev.Lev02_&GRADE. (DROP= _TYPE_ _FREQ_);
   SET Raw&TestAbrev.Lev02temp_&GRADE.;
   WHERE _FREQ_ GE &MinNumStudnts.;
RUN;
PROC SORT DATA=Raw&TestAbrev.Lev02_&GRADE.;
   BY SCHL1D;  RUN;
PROC SORT DATA=Raw&TestAbrev.Lev01_&GRADE.;
   BY SCHL1D;  RUN;
DATA Raw&TestAbrev.Data&GRADE.;
   MERGE Raw&TestAbrev.Lev02_&GRADE. (IN=valid) Raw&TestAbrev.Lev01_&GRADE.;
   BY SCHL1D;
   IF valid;
RUN;
PROC SORT DATA= Raw&TestAbrev.Data&GRADE.;
   BY SCHL1D;
RUN;
DATA &TestAbrev.&GRADE.; SET Raw&TestAbrev.Data&GRADE. (KEEP=SCHL1D GRAD1D &test.);
   WHERE GRAD1D = &GRADE.;
RUN;
%MEND Sets;
&InvokeSets.;


/**********************************PARAMETER ESTIMATION***********************************/


%MACRO ICC (GRADE= , Test= ,TestAbrev= , SampSize= ,tdist= ,cov= ,CovLabel= );

    /*Calculates harmonic mean (required if your plan is to include ALL students in the sampled schools in your future study)*/
DATA &TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.;
   SET &TestAbrev.&GRADE.;
RUN;
PROC SQL;
  CREATE TABLE HarMtemp&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel. AS
```

```
     SELECT COUNT(&test.) AS N, 1/(Calculated N) AS invN
    FROM &TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel. group BY SCHL1D;
QUIT;
PROC SQL;
   CREATE TABLE HarMean&TestAbrev.&GRADE.&SampSize. AS
     SELECT (COUNT(invN)/sum(invN)) AS Harmean, (sum(N)/COUNT(N)) AS Aritmean, COUNT(N) AS CountSchls, "&SampSize." AS SampSize, "&Test."
     AS Test, "&Grade." AS Grade,"&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel." AS TempID
    FROM HarMtemp&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.;
QUIT;

    /*Estimates variance components, using PROC Mixed (REML), used for calculating ICCs*/
ODS LISTING CLOSE;
ODS OUTPUT  CovParms=&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.tempa;
PROC MIXED DATA=&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel. METHOD=reml covtest cl ;
   CLASS SCHL1D;
   MODEL &test. = /  DDFM=bw cl;
   RANDOM INTERCEPT/ SUBJECT=SCHL1D;
RUN;
ODS OUTPUT  CLOSE;
ODS LISTING;

    /*Uses estimates of the variance components to calculate ICC, ICC confidence limits, and MDESs.*/
DATA &TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.tempb;
   SET &TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.tempa; KEEP CovParm Estimate StdErr Lower Upper;
RUN;

DATA &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_tau (KEEP=Tau TauSE TauLwr TauUpr TempID)
      &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_sig (KEEP=Sigma SigmSE SigmLwr SigmUpr TempID);
   LENGTH TempID $ 32;
   SET &TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.tempb;
   IF CovParm = "Intercept" THEN DO;
      TempID = "&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.";
      Tau = Estimate;
      TauSE = StdErr;
      TauLwr = Lower;
      TauUpr = Upper;
     OUTPUT  &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_tau;
   END;
   ELSE IF CovParm= "Residual" THEN DO;
      TempID = "&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.";
      Sigma = Estimate;
      SigmSE = StdErr;
      SigmLwr = Lower;
      SigmUpr = Upper;
     OUTPUT  &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_sig;
   END;
RUN;


DATA &TestAbrev.&GRADE.&SampSize.temp&CovLabel.;
```

```
   LENGTH ID $ 32;
   MERGE &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_tau  &TestAbrev.&GRADE.r_&SampSize.Cov&CovLabel.t_C_sig
HarMean&TestAbrev.&GRADE.&SampSize.;
   BY TempID;
   ID = "&TestAbrev.&GRADE.ran&SampSize.Cov&CovLabel.";
   Cov = &cov.;
   IF Cov > 0 THEN DFminus = 2; ELSE DFminus = 1;
   ICC=Tau/(Tau+Sigma);
   Var_ICC = (2*((1-ICC)**2)*((1+(&NL1.-1)*ICC)**2))/(&NL1.*(&NL1.-1)*&SampSize.);
   SE_ICC  = sqrt(Var_ICC);
   ICC_Low = ICC-(&tdist.*sqrt(Var_ICC)); If ICC_Low LT 0 THEN ICC_Low = 0;
   ICC_Hig = ICC+(&tdist.*sqrt(Var_ICC)); If ICC_Hig GT 1 THEN ICC_Hig = 1;
   CI_Range = ICC_Hig - ICC_Low;
 RUN;
DATA &TestAbrev.&GRADE.&SampSize.&CovLabel.;
   SET &TestAbrev.&GRADE.&SampSize.temp&CovLabel.;
   tdist = TINV(1-.05/2,sampsize-1);
   tbeta = TINV(.80, sampsize-DFminus);
   M = tdist + tbeta;
   P = &PcntTreatGrp.;
   MDES = M*(1/(sqrt(P*(1-P))))*(SQRT(((ICC*(1-&cov.))/SampSize)+((1-ICC)/(SampSize*&NL1.))));
   MDES_uppr = M*(1/(sqrt(P*(1-P))))*(SQRT(((ICC_Hig*(1-&cov.))/SampSize)+((1-ICC_Hig)/(SampSize*&NL1.))));
   RUN;
QUIT;
%MEND ICC;
&InvokeICCs.;


/**********************************MERGING All Grades' ICCs************************************

    Places the subsets (as specified by the user; i.e., by grade level, test, and number of schools) in a file for printing
    (The PUT statement prints, in the log, which datasets are being merged)*/

DATA All;
   LENGTH test $ 8;
   LENGTH SampSize $10;
   MERGE &InvokeMerg.;
   BY ID;
   %PUT ****The merged sets are &InvokeMerg.*****;
RUN;
PROC SORT DATA=All;
   BY test ID;
RUN;
```

```
/*****************************************PRINTING*********************************************
    Note. The user may wish to add ODS output here [such as ODS CSV file = (your directory and file
    name);] to output to a CSV file readable in Excel*/

OPTIONS pageno=1 linesize=120 pagesize=54 nocenter;
ODS HTML close;
ODS HTML;

PROC PRINT DATA=All noobs label;
   BY test;
   VAR  Grade test SampSize CountSchls Cov  ICC SE_ICC ICC_Low ICC_Hig  MDES MDES_Uppr;
   FORMAT   ICC
            SE_ICC
            ICC_Low
            ICC_Hig
            Cov       4.3
            MDES
            MDES_Uppr 5.3
            HarMean
            AritMean 10.2
            Var_ICC  10.4
            CountSchls 10.0
            SampSize $10.0
            ;
   LABEL    ID      = "Analysis ID"
            ICC     = "ICC"
            SE_ICC  = "Standard error of ICC"
            ICC_Low = "Lower ICC"
            ICC_Hig = "Upper ICC"
            CountSchls= "Total number of schools"
            SampSize= "Sampled number of schools"
            HarMean = "HarMean N students per school"
            AritMean= "AritMean N students per school"
            MDES_Uppr="  Upper MDES"
            Cov      =" Specified Covariance"
            ;
     TITLE1 "Unconditional Intraclass Correlation Confidence Limits and MDES and Upper MDES Estimates";
     TITLE2 "  (&Note.)";
   FOOTNOTE1 "Notes:";
   FOOTNOTE2 "*The 'Sampled number of schools' is the number you plan to randomly sample from this population.";
   FOOTNOTE3 "*The 'Total number of schools' is the original population (or larger sample) from which you are sampling your schools.";
   FOOTNOTE4 "The 'Specified Covariance' is the cluster level (school level) covariance (r-squared) you specified (used in calculating
              MDES).";
   FOOTNOTE5 "All schools used from your original population have a min of &MinNumStudnts. students in the specified grade.";
   FOOTNOTE6 "You specified that the number of students per school you plan to sample is &NL1..";
RUN;
   TITLE; FOOTNOTE;

QUIT;
```