Paper 217-25

Robust Effect Size Estimates and Meta-Analytic Tests of Homogeneity

Kristine Y. Hogarty and Jeffrey D. Kromrey

Department of Educational Measurement and Research, University of South Florida

ABSTRACT

This paper presents a computer program that calculates two effect size estimates (gamma and trimmed-d) that are robust to violations of the assumptions of population normality and homogeneity of variance. Because of their robustness properties, these indices are frequently more appropriate than the conventional parametric indices such as Cohen's **d** or Hedges' **g**. Additionally, a program that uses these effect size indices in robust permutation tests for between-class homogeneity (a test frequently used in meta-analysis) is introduced. The paper provides a demonstration of the SAS/IML® code and examples of the application of the code in simulation studies.

INTRODUCTION

Effect size estimates are frequently used in planning research, determining the practical significance of research results, and comparing results across studies (Fern & Monroe, 1996). The latest publication manual of the American Psychological Association (APA, 1994) strongly encourages the reporting of effect sizes as research results. With the increasing use of effect sizes and the increasing recognition of meta-analysis as an important research tool, a concomitant recognition of the sensitivity of effect sizes and meta-analytic tests to violations of assumptions has been evidenced (Hedges & Olkin, 1985; Harwell, 1997; Hogarty & Kromrey, 1999). Kraemer and Andrews (1982) argued that the use of parametric indices of effect size, such as those proposed by Glass (1976) and Cohen (1988), may lead to biased conclusions if the underlying assumptions (i.e., normality, homogeneity of variance) are violated.

In an attempt to overcome some of the problems encountered when using typical indices of effect size (e.g., Cohen's **d** or Hedges' **g**) alternatives to these parametric estimators have been suggested. Hedges and Olkin (1985) proposed a nonparametric effect size estimate, the γ_1^* index based on proportions of overlap between two samples. Alternately, the use of robust estimates of means and variances in primary studies has also been suggested (Yuen, 1974; Hedges & Olkin, 1985). The approach recommended by Yuen (1974) is appealing because of its computational ease and strong theoretical properties.

In addition to sensitivity of parametric effect size indices, an increasing concern has been expressed about the sensitivity of meta-analytic tests of homogeneity (e.g., Hedges' Q test). Harwell (1997) provided evidence that the Q test did not control Type I error rates when the assumptions of population normality and homogeneity of variance were violated. Kromrey and Hogarty (1999) demonstrated that permutation tests applied in meta-analytic contexts provided exceptional Type I error control even under extreme violations of these assumptions. As a follow-up, Hogarty and Kromrey (1999) provided estimates of statistical power for randomization tests using robust effect size estimates.

EFFECT SIZE INDICES EXAMINED

A variety of effect size indices are available to researchers (Rosenthal, 1994). The most commonly applied are those that represent standardized mean differences (i.e., Cohen's **d** and Hedges' **g**). Because these standardized mean difference effect size estimates may be sensitive to violations of the assumptions of normality and homogeneity of variance, some nonparametric indices have been proposed. One of these, which is applicable to describing the effect size of the difference between two independent groups is the index described by Hedges and Olkin (1985).

$$\gamma_1^* = \Phi^{-1}(q^*)$$

where q* is the sample proportion of scores in one group that are less than the median score of the other group, and Φ^{-1} is the inverse of the standard normal cumulative distribution function.

The γ_1^* index is, therefore, the normal deviate that corresponds to the proportion q*. In practice, if the observed q* = 0 or 1 (for which the inverse transformation yields negative or positive infinity) the proportion is replaced with 1/(n + 1) or n/(n + 1), respectively.

An alternative to nonparametric effect size estimators is the use of robust estimates of means and variances (e.g., trimmed means and Winsorized variances; Yuen, 1974) in primary studies. The use of trimmed means and Winsorized variances in the construction of robust effect size indices was suggested by Hedges and Olkin (1985). The trimmed mean for a sample of scores is obtained by dropping the highest and lowest *k* scores from the sample before the mean is computed.

$$\bar{X}_t = \frac{X_{k+1} + X_{k+2} + \ldots + X_{n-k}}{n-2k}$$

Similarly the Winsorized variance, a robust variance estimate, is the sample variance computed by "replacing" the lowest *k* values by the *(k + 1)th* value, and "replacing" the highest *k* values by the *(n - k)th* value.

$$S_w^2 = \left[\frac{1}{n-2k}\right]\Big[(k+1)(X_{k+1} - \bar{X}_w)^2 + (X_{k+2} - \bar{X}_w)^2 + \ldots + (k+1)(X_{n-k} - \bar{X}_w)^2\Big]$$

where $\bar{X}_w$ is the Winsorized mean:

$$\bar{X}_w = \frac{(k+1)(X_{k+1}) + X_{k+2} + \ldots + (k+1)(X_{n-k})}{n}$$

META-ANALYTIC RANDOMIZATION TESTS

Meta-analytic tests for differences in effect size across classes of studies (e.g., studies conducted with male participants vs. those conducted with female participants) may be obtained using a randomization testing procedure (Noreen, 1989; Edgington, 1987).

In a randomization strategy, the assignment of studies to each of the classes is randomly permuted. For each randomization of the data, the difference in weighted mean effect sizes between the classes is calculated. The set of mean differences provides an empirical sampling distribution within which the actual mean difference observed in the meta-analysis may be compared, thus providing a probability under the null hypothesis of no differences between classes in mean effect size.

For meta-analyses consisting of relatively few studies, an exact permutation test may be conducted by calculating all possible permutations of the effect sizes. However, for meta-analyses consisting of larger numbers of studies, these computations become prohibitive and approximate randomization tests are used (in which 5,000 or 10,000 randomly sampled permutations are used to construct the sampling distribution).

AN EXAMPLE

A SAS macro was designed to calculate the effect size indices γ_1^* and trimmed-d using raw data obtained from two groups of observations. These statistics will be described in reference to the set of data presented in Table 1. These data consist of nine observations obtained in a control group and nine independent observations obtained in an experimental group. The sample means for these data are identical (16.22) in the two groups, although the variance of the control group (S^2 = 26.19) is more than three times that of the experimental group (S^2 = 7.69). Because the means are identical, both Cohen's **d** and Hedges **g** effect sizes are zero.

For the computation of the γ_1^* effect size, the sample medians are computed (16.0 for the control group and 17.0 for the experimental group). Using the control group median as the reference point, 4 of the 9 observations (or 0.444) in the experimental group are lower than the control group median. This proportion of a standard normal curve corresponds to a z-score of approximately –0.14, which is the robust effect size estimate.

The trimmed-d effect size is obtained by dropping the highest and lowest observation from each group and calculating the resulting trimmed means (15.57 and 16.43 for the control and experimental groups, respectively). Similarly, the Winsorized variances are obtained by replacing these most extreme scores by the next closest scores observed in each group. For these data, the Winsorized variances are 12.00 for the control group and 4.60 for the experimental group. The resulting trimmed-d effect size (the ratio of the difference in trimmed means to the square root of the pooled Winsorized variance) is approximately -0.30.

Table 1
Sample of Two Groups' Posttest Scores

Control Group	Experimental Group
10	11
12	14
12	15
15	15
16	17
16	17
18	18
20	19
27	20

MACRO EFF_SIZE

The macro EFF_SIZE calculates the two robust effect size estimates. Arguments to the macro are the name of the SAS dataset containing the data, the variable that contains scores on the dependent variable, and the variable that indicates group membership. As written, the macro expects this grouping variable to take the values of 1 and 2, which represent observations from the control and experimental groups, respectively.

The data are passed to PROC IML and calculations are accomplished in three subroutines. The subroutine BUBBLE, used to obtain the sample median, sorts the data. The subroutine GAMMA calculates the γ_1^* effect size and the subroutine TRIMMIT computes trimmed means and Winsorized variances. The output of the macro (obtained via the FILE PRINT command) is presented in Table 2.

```
%macro eff_size(testdata,grpvar,dvscore);
 proc iml;
  use &testdata;
  read all var {&dvscore} where
  (&grpvar=1) into A;
  read all var {&dvscore} where
  (&grpvar=2) into B;
 NA=NROW(A);
 NB=NROW(B);
 score={&dvscore};
 group={&grpvar};

 * +----------------------------------+
     Bubble sort
   +----------------------------------+;
START BUBBLE(x,n,c);
 do i = 1 to n;
  do j = 1 to n-1;
    if x[J,C] > x[J+1,C] then do;
       temp = x[J+1,];
       x[J+1,] = x[J,];
       x[J,] = temp;
    end;
  end;
end;
FINISH;

START GAMMA(A,B,NA,NB,expmedn,ctlmedn,
            gamma1);
 score={&dvscore};
 group={&grpvar};

* Sort the data in each group and find
  the median of group 1 to use in the
  gamma-star 1 effect size;

 run bubble(A,NA,1);
 run bubble(B,NB,1);

 IF 0.5*NA = ROUND(0.5*NA) THEN even=1;
 IF 0.5*NA ^= ROUND(0.5*NA) THEN even=0;
 IF even=0 THEN DO;
      ctlmedn= A[0.5*NA + 0.5,1];
 END;

 IF even=1 THEN DO;
      ctlmedn= 0.5*(A[(0.5*NA),1] +
               A[(0.5*NA   + 1),1]);
 END;

 IF 0.5*NB = ROUND(0.5*NB) THEN even=1;
 IF 0.5*NB ^= ROUND(0.5*NB) THEN even=0;
 IF even=0 THEN DO;
      expmedn= B[0.5*NB + 0.5,1];
 END;

 IF even=1 THEN DO;
      expmedn= 0.5*(B[(0.5*NB),1] +
               B[(0.5*NB + 1),1]);
 END;

* Calculate the gamma effect size;
    Countlss=0;
 do g = 1 to NB;
     if B[g,] < ctlmedn then countlss =
     countlss + 1;
 end;
 if (countlss > 0 & countlss < NB) then
   gamma1 = probit(countlss/NB);
 if countlss=0 then gamma1 =
   probit(1/(NB+1));
 if countlss=NB then gamma1 =
   probit(countlss/(NB+1));
 FINISH;

START TRIMMIT (XX,trimpct,trim,
      T_mean,W_var);
  n_obs = NROW(XX);
  trim=round((trimpct/100)#n_obs + 0.5);
  XT = J(n_obs - 2*trim,1,0);
   do t = 1 to (n_obs - 2*trim);
     XT[t] = XX[t+trim];
   end;
  T_mean = 0;
  W_mean = 0;
   do t = 1 to (n_obs - 2*trim);
   if (t  = 1 | t  = n_obs - 2*trim) then
    wt = trim + 1;
   if (t ^= 1 & t ^= n_obs - 2*trim) then
    wt = 1;
  W_mean = W_mean + wt*XT[t];
  T_mean = T_mean + XT[t];
   end;
   W_mean = W_mean/n_obs;
   T_mean = T_mean/(n_obs - 2*trim);

 * Compute Winsorized variance;
   W_var = 0;
   do t = 1 to (n_obs - 2*trim);
   if (t  = 1 | t  = n_obs - 2*trim) then
    wt = trim + 1;
   if (t ^= 1 & t ^= n_obs - 2*trim) then
    wt = 1;
  W_var = W_var + wt*(XT[t] - W_mean)**2;
   end;
  W_var = W_var/(n_obs - 2*trim);
 FINISH;

* Calculate the trimmed effect size;
 run trimmit(A,10,t_count1,T_mean1,
   W_var1);
 run trimmit(B,10,t_count2,T_mean2,
   W_var2);
 trim_es = (T_mean1 - T_mean2) /
   SQRT((W_var1*(NA - 2*t_count1) +
   W_var2*(NB - 2*t_count2)) /
   ((NA - 2*t_count1) +
    (NB - 2*t_count2)));
run gamma(A,B,NA,NB,expmedn,ctlmedn,
          gamma1);

file print;
 put  @28 'Sample Robust Effect Sizes'/
      @10 ________________________'//
   @10  'Dependent Variable' @40 SCORE //
 @10  'Between Class Variable' @40 GROUP //
 @57  'Effect'/
 @34  'Group  1' @45 'Group  2' @58 'Size'/
 @34  '____' @45 '____' @56 '________'//
 @10  'Sample Sizes' @34 NA 8. @45 NB 8. //
 @10  'Trimmed Means' @34 T_mean1 8.5
 @45   T_mean2 8.5 @56 trim_es 8.5 /
 @10  'Winsorized Variance' @34 W_var1 8.5
 @45   W_var2 8.5 //
 @10  'Median' @34 ctlmedn 8.5
 @45   expmedn 8.5 @56 gamma1 8.5//
 @10  '__________________________________'/;
quit;

%mend eff_size;
```

MACRO PERMUTE

The macro PERMUTE calculates approximate permutation tests for differences in mean effect sizes between two groups of studies. The arguments to the macro are (a) the name of the SAS dataset containing effect sizes from individual studies, (b) the variables that contain the sample effect sizes (both the trimmed-d and the γ_1^* effect sizes), and (c) the variable that indicates the group in which the study has been classified (as written, the macro expects this variable to take the values of 1 and 2).

The data are passed to PROC IML and calculations are accomplished in two subroutines. The subroutine RESAMP produces a random permuation of the data into two groups. The subroutine PERMTEST calculates the observed mean difference in effect sizes between the two classes of studies, then calls the RESAMP subroutine 5000 times. For each permutation of the data, the observed mean difference in effect size is compared to that observed in the actual data. The proportion of permutations with mean differences smaller than the difference observed in the actual samples provides the probability estimate associated with equality of population effect sizes. The output of the macro (obtained via the FILE PRINT command) is presented in Table 3.

```
%macro permute(testdata,gam_var,trim_var,
               grp_var);
proc iml;
* +-----------------------------------+
   Direct resampling for randomization
  +-----------------------------------+;
START RESAMP(x);
n=Nrow(x);
allnbut=n-1;
  do i = 1 to allnbut;
    ranrow = round(uniform(0)*(n - i +
             0.999)+0.5);
    if i = 1 then do;
       newm = x[ranrow,];
    end;
    if i > 1 then do;
       newm = newm//x[ranrow,];
    end;
    if ranrow > 1 then do;
        if ranrow < (n-(i-1)) then
          x = x[1:ranrow-1,]//
              x[ranrow+1:n-(i-1),];
        if ranrow = n-(i-1) then
          x=x[1:(n-i),];
    end;
    if ranrow = 1 then x = x[2:n-(i-1),];
  end;
  newm = newm//x;
  x = newm;
FINISH;

START PERMTEST(gam_vec1,gam_vec2,
        trm_vec1, trm_vec2, n_perms);
  k1 = nrow(gam_vec1);
  k2 = nrow(gam_vec2);
  k_total= k1 + k2;
  eff_mtx = (gam_vec1//gam_vec2) ||
            (trm_vec1//trm_vec2);
  mn_gam1 = sum(gam_vec1)/k1;
  mn_gam2 = sum(gam_vec2)/k2;
  mn_trm1 = sum(trm_vec1)/k1;
  mn_trm2 = sum(trm_vec2)/k2;
  gam_diff = mn_gam1 - mn_gam2;
  trm_diff = mn_trm1 - mn_trm2;
  prob_gam=0;
  prob_trm=0;
  gamt1=0;
  trmt1=0;
  perm = 0;
do i=1 to n_perms;
   run resamp (eff_mtx);
   gamt1 = eff_mtx[1:k1,1];
   gamt1 = sum(gamt1)/k1;
   gamt2 = eff_mtx[k1+1:k_total,1];
   gamt2 = sum(gamt2)/k2;
   trmt1 = eff_mtx[1:k1,2];
   trmt1 = sum(trmt1)/k1;
   trmt2 = eff_mtx[k1+1:k_total,2];
   trmt2 = sum(trmt2)/k2;
   if abs(gamt1 - gamt2) < abs(gam_diff)
       then prob_gam = prob_gam+1;
   if abs(trmt1 - trmt2) < abs(trm_diff)
       then prob_trm = prob_trm + 1;
   perm=perm+1;
   free gamt1 gamt2 trmt1 trmt2;
end;
prob_gam = 1 - (prob_gam/perm);
prob_trm = 1 - (prob_trm/perm);
file print;
put @1 'Approximate Randomization Test of
Homogeneity of Effect Sizes' /
 @1 '_______________________________
______________________________' /
 @1 'N of Permutations' @30 perm 8. //
 @31 'Group 1' @41 'Group 2'
 @51 'Prob > |ES|' /
 @31 '_______' @41 '_______'
 @51 '__________' /
 @1 'N of Studies'  @30 k1 8.
 @40 k2 8. /
 @1 'Mean Gamma'    @30 mn_gam1 best8.
 @40 mn_gam2 best8.
 @53 prob_gam best8. /
 @1 'Mean Trimmed-d' @30 mn_trm1 best8.
 @40 mn_trm2 best8.
 @53 prob_trm best8. //
@1 '________________________________
_____________________________' ;
FINISH;
 use &testdata;
 read all var {&gam_var} where
         (&grp_var=1) into gam_vec1;
 read all var {&gam_var} where
         (&grp_var=2) into gam_vec2;
 read all var {&trim_var} where
         (&grp_var=1) into trm_vec1;
 read all var {&trim_var} where
         (&grp_var=2) into trm_vec2;

 run permtest (gam_vec1,gam_vec2,
       trm_vec1,trm_vec2, 5000);
quit;
%mend permute;
```

OUTPUT

The output from the macro EFF_SIZE is presented in Table 2. In addition to the name of the dependent variable and the name of the grouping variable, the output includes the sample sizes, trimmed means, Winsorized variances and sample medians. Finally, the two effect size estimates are provided.

Table 2

Sample Robust Effect Sizes

Dependent Variable SCORE

Between Class Variable GROUP

	Group 1	Group 2	Effect Size
Sample sizes	9	9	
Trimmed Means	15.57143	16.42857	-0.29749
Winsorized Variance	12.00000	4.60317	
Median	16.00000	17.00000	-0.13971

The output from the macro PERMUTE is presented in Table 3. The printed output includes the number of permutations used in the tests, the number of studies in each group and the mean value of each effect size observed in each of the two groups. Finally, the two-tailed probability under the null hypothesis of equal population effect sizes is reported.

Table 3

Approximate Randomization Test of Homogeneity of Effect Sizes

N of Permutations 5000

	Group 1	Group 2	Prob>|ES|
N of Studies	4	6	
Mean Gamma	0.228	0.6345	0.057
Mean Trimmed-d	0.73625	0.516	0.153

CONCLUSION

In many research applications, the assumptions of parametric statistical tests are violated. Because parametric effect size indices are based on the same assumptions, their use with such data may be misleading, and their use in meta-analysis may lead to inflated Type I error rates. The robust effect size indices combined with meta-analytic permutation tests have been shown to evidence superior Type I error control and statistical power under conditions in which the assumptions of population normality and homogeneity of variance are violated in the primary studies (Hogarty & Kromrey, 1999; Kromrey & Hogarty, 1999). However, the calculation of these effect size estimates requires more detailed information about observed data than is typically reported in research publications.

The macros are provided to facilitate researchers' use of these robust statistical methods. Although the macros are presented for the analysis of two groups, they are easily modified to incorporate data with more than two groups. Similarly, the PERMUTE macro is presented in a form that provides approximate permutation tests. A version of this macro that computes exact permutation tests is available from the authors.

REFERENCES

American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Edgington, E. S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.

Fern, E. F. & Monroe, K. B. (1996). Effect size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89-105.

Glass, G. V. (1976). Primary, secondary, and meta-analysis research. *Educational Researcher*, *5*, 3-8.

Harwell, M. (1997). An empirical study of Hedges' Homogeneity Test. *Psychological Methods*, *2*, 219-231.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hogarty, K. Y. & Kromrey, J. D. (1999, August). *Tradtional and robust effect size estimates: Power and Type I error control in meta-analytic tests of homogeneity*. Paper presented at the Joint Statistical Meetings, Baltimore.

Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, *91*, 404-412.

Kromrey, J. D. & Hogarty, K. Y. (1999, April). *Traditional and robust effect size estimates: An empirical comparison in meta-analytic tests of homogeneity*. Paper presented at the annual meeting of the American Educational Research Assocation, Montreal.

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses: An introduction*. New York: Wiley.

CONTACT INFORMATION

The authors can be contacted at the University of South Florida, Department of Educational Measurement and Research, EDU 162, 4202 E. Fowler Avenue, Tampa, FL 33620. They may be contacted by telephone at (813) 974-3220 or by electronic mail at khogarty@luna.cas.usf.edu or kromrey@typhoon.coedu.usf.edu