The Stata Journal (2024)
24, Number 3, pp. 503–513 DOI: 10.1177/1536867X241276113

miesize: Effect-size calculation in imputed data

Paul A. Tiffin
Hull York Medical School
University of York
York, UK
paul.tiffin@york.ac.uk

Abstract. In this article, I describe the `miesize` command for the calculation of effect sizes in imputed data. There may be situations where an effect size needs to be estimated for an intervention, an exposure, or a group membership variable but data on the independent or dependent variable are missing. Such missing data are commonly dealt with by multiply imputing plausible values. However, in this circumstance, the estimated effect size and associated standard errors will need to be pooled and estimated from the imputed dataset. The `miesize` command automates this process and calculates effect sizes for a binary variable from multiply imputed data in wide format. The estimates and standard errors (used to calculate the confidence intervals) are recombined using Rubin's (1987, *Multiple Imputation for Nonresponse in Surveys* [Wiley]) rules. These rules are applied such that the average point estimate for the effect size is calculated from the imputed datasets. The pooled standard error, and hence confidence intervals, is calculated to account for both the variance between the imputed datasets and the variance within them. Pooled effect sizes and confidence intervals for Cohen's (1988, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. [Lawrence Erlbaum]) d, Hedges's (1981, *Journal of Educational Statistics* 6: 107–128) g, and Glass's (Smith and Glass, 1977, *American Psychologist* 32: 752–760) delta are provided by `miesize`.

Keywords: st0755, miesize, effect size, imputation, Rubin's rules, Cohen's d, Hedges's g, Glass's delta

1 Introduction

The term "effect size" usually refers to the magnitude, and direction, of an association between two variables or the mean difference between two groups. Statistical tests that produce p-values indicate the probability that the observed result, or one more extreme, was due to chance alone. In contrast, effect sizes indicate the magnitude of the association between variables, or mean group difference. Such estimation is intended to help interpret the results and understand what they may mean in practical terms. For example, in a healthcare situation, an experimental treatment may result in a statistically significant difference in a symptom outcome between the experimental and control group. However, the treatment is unlikely to be useful in practice unless the *magnitude* of the difference is clinically meaningful—that is, one that makes a difference to the average quality of life of patients.

In relation to estimating a mean difference for a continuous outcome across groups, these methods are sometimes referred to as the "d" family. The average group difference

described by an effect size could be applied to experimental data. In this situation, the mean value of an outcome of interest between a group of individuals randomized to either an experimental intervention or a control condition is compared and contrasted. In observational data, the mean value can refer to a mean intergroup difference in relation to a sociodemographic or other characteristic.

Effect sizes, relating to a mean group difference, can be described in the original metric of an outcome measure. However, these may not be easily interpretable. For example, the outcome may relate to an attitudinal questionnaire, with responses scored and summed according to a Likert scale. Some experts may understand what a difference of "five points" may mean in this situation, but many others would not. Thus, the term "effect size" often refers to such differences quantified in a standardized metric. The d family of estimators provides an effect size that is standardized according to the standard deviation (SD) (that is, square root of the variance) of the outcome of interest. All the d family of estimators assume that the outcome of interest is normally distributed, and thus, departure from this can introduce bias (Grissom and Kim 2001).

One of the most commonly used metrics of effect size is Cohen's (1988) d. For Cohen's d, standardization is performed according to the pooled SD for both groups being compared (see below). In general, the d family communicates effect size as the scaled difference between the means, divided by the SD of the outcome of interest. Here we see in the following equation that Cohen's d is given by the difference in the means of the outcome between the two groups ($\overline{x}_1$ and $\overline{x}_2$) divided by the pooled SD (s^*):

$$d = \frac{(\overline{x}_1 - \overline{x}_2)}{s^*}$$

The pooled SD (s^*) is calculated as

$$s^* = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

where n_1 and n_2 refer to the number of observations in each of the two comparison groups, respectively. Here s_1 and s_2 are the SDs of the outcome for each respective group.

However, Hedges showed that Cohen's d can be biased, especially when there are relatively small numbers of observations (for example, $N < 20$), and proposed a correction factor could be applied to d (Hedges 1981; Hedges and Olkin 1985). Let d represent Cohen's d and m the summed total number of observations across groups (that is, $m = n_1 + n_2$). Where Γ represents the gamma function, Hedges's g is calculated as $g = d \times c(m)$, where

$$c\,(m) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{\frac{m}{2}}\,\Gamma\left(\frac{m-1}{2}\right)}$$

Though not implemented in Stata, Hedges also provided a simplified version of the correction factor. This is slightly biased but computationally easier to calculate. Here d.f. represents the degrees of freedom, given by d.f. $= n_1 + n_2 - 2$, for an independent

group design (Borenstein et al. 2021). Once again, d represents Cohen's d, and this approximation (g^*) is calculated as

$$g^* = d \times \left(1 - \frac{3}{4\,(\text{d.f.}) - 1}\right)$$

Both Cohen's d and Hedges's g use the pooled variance of the outcome. In contrast Glass's delta (Δ) provides two effect sizes, Δ_1 and Δ_2, relating to the variance in the respective groups of observations. Where $\overline{x}_1$ and $\overline{x}_2$ once again represent the group means of the continuous outcome, and $s_1$and s_2 represent the group specific SDs for the outcome, Glass's Δs are calculated as

$$\Delta_1 = \frac{\overline{x}_1 - \overline{x}_2}{s_1}$$
$$\Delta_2 = \frac{\overline{x}_1 - \overline{x}_2}{s_2}$$

Originally, this approach was intended to be applied to experimental data where Δ_1 related to the control group (Smith and Glass 1977). For experimental data, usually the SD for the control group is used when calculating Δ. However, for observational data, the choice of which of the group's SD to use appears arbitrary, while for dummy-coded group variables (one or zero/absent or present), the group with "zero" status could be the one to have the SD used in calculating Δ. However, in the absence of a compelling reason to consistently report Δ based on one or the other group's SD, it has been recommended that both effect-size estimates (Δs) be reported for observational data (Kline 2013).

Irrespective of the method used, there are well-known rules of thumb for interpreting effect sizes. That is, effect sizes of 0.2 to 0.5 are usually regarded as "small/modest", 0.5 to 0.8 as "medium sized", and 0.8 or above as "large sized". However, these interpretations will need to be contextualized. For example, in healthcare it is for stakeholders and experts (patients and clinicians) to agree on what magnitude of difference may be considered to constitute a clinically meaningful effect size. Moreover, from an economic perspective, cost effectiveness of an intervention may not be achieved even with an effect size that is classed as "large" in this context. Thus, it is important to understand the meaning and implications of an effect size within its specific substantive context.

Missing data may be present in both experimental design and observational studies. Analyzing data using listwise deletion may give biased results. It is also wasteful of the information available in the remaining values in the variables that are present in observations with one or more missing values. Thus, best practice when analyzing data with missing values is to use multiple imputation (van Ginkel et al. 2020; Sterne et al. 2009). This involves drawing plausible values for those missing from conditional probability distributions. These distributions are shaped by the relationship observed between the variables for which the values are nonmissing. These values are imputed for multiple datasets (usually five or more), and the results recombined. Such results are unbiased if the data are missing completely at random (that is, because of chance

only) or missing at random (missing values depend on the observed values of variables) according to Rubin's (1987) missing data mechanism classification. Indeed, even if the data are not missing at random (the missing values may depend on unobserved variable values), the results from multiply imputed data may be less biased than that derived from listwise deletion (van Ginkel et al. 2020). When recombining the results from multiply imputed data, one must account for the uncertainty introduced by the imputation process when deriving standard errors (SEs). Stata currently offers a wide range of analyses that work with multiply imputed data. These are invoked by using the command after the prefix `mi estimate`, for example, `mi estimate: regress` ....

Rubin (1987) provided rules for how means and SEs (and hence confidence intervals) could be combined from results derived from multiply imputed datasets. Known aptly enough as "Rubin's rules", these are as follows.

To pool an effect estimate, such as the pooled mean difference ($\overline{\theta}$), one uses the following formula. Here, on this occasion, m represents the number of imputed datasets used, and i denotes the ith dataset:

$$\overline{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m}\theta_i\right)$$

This is effectively the mean of the mean differences calculated across the m imputed datasets.

Pooling the SE for such estimates is more complicated and must account for the sampling variance both within and between the imputed datasets. The "within imputation variance" is the average of the mean of the within variance estimate. In effect, this is the squared SE calculated for each imputed dataset. This value reflects the sampling variance in each of the datasets generated by the multiple imputation process. Thus, as expected, this value will be relatively large in small samples and more modest in larger samples (Heymans and Eekhout 2019). It is given by the following equation, where V_W is the within imputation variance:

$$V_W = \frac{1}{m}\sum_{i=1}^{m}\mathrm{SE}_i^2$$

The between imputation variance is intended to reflect the additional variance due to the uncertainty relating to the value of the imputed data. This is estimated by calculating the variance of the parameter of interest estimated over all the imputed datasets. This formula is the same as that ordinarily used to calculate the variance in a given sample. The value is large when the missing data are extensive and relatively smaller with fewer missing data. Here V_B is the between imputation variance, $\overline{\theta}$ is the overall pooled estimate for the parameter of interest, and θ_i is the parameter of interest estimated in each of the m imputed datasets:

$$V_B = \frac{\sum_{i=1}^{m}\left(\theta_i - \overline{\theta}\right)^2}{m-1}$$

To date, Stata does not include a command to estimate effect sizes from multiply imputed data. This function may be useful given that missing data are regularly encountered in experimental and observational data where it may be desirable to calculate effect sizes.

2 The miesize command

2.1 Syntax

`miesize` *varname* [*if*] [*in*], `by(`*groupvar*`)` [`glass countdown level(`#`)`]

varname is the outcome variable of interest. *groupvar* is a variable that defines the two groups that `miesize` will use to estimate the effect sizes.

The command returns a range of results in `r()`. Do not confuse the `by()` option with the `by` prefix; you can specify only the former in `miesize`.

2.2 Options

`by(`*groupvar*`)` specifies the *groupvar* that defines the two groups that `miesize` will use to estimate the effect sizes. `by()` is required.

`glass` reports Glass's Δ (Smith and Glass 1977) using each group's SD.

`countdown` specifies that a countdown of analysis steps remaining be displayed.

`level(`#`)` specifies the confidence level to be reported. This can be set between 10 and 99.99%. The default is `level(95)`.

2.3 Stored results

`miesize` stores the following results in `r()`:

If the `glass` option is not specified:

Macros

`r(pooled_se_d)`	pooled SE for Cohen's d
`r(pt_est_d)`	pooled point estimate for Cohen's d
`r(ub_d)`	upper confidence limit for the estimate of Cohen's d
`r(lb_d)`	lower confidence limit for the estimate of Cohen's d
`r(pooled_se_g)`	pooled SE for Hedges's g
`r(pt_est_g)`	pooled point estimate for Hedges's g
`r(ub_g)`	upper confidence limit for the estimate of Hedges's g
`r(lb_g)`	lower confidence limit for the estimate of Hedges's g

If the `glass` option is specified:

Macros

`r(pooled_se_g1)`	pooled SE for Glass's Δ for group 1
`r(pt_est_g1)`	pooled point estimate for Glass's Δ for group 1
`r(ub_g1)`	upper confidence limit for the estimate of Glass's Δ for group 1
`r(lb_g1)`	lower confidence limit for the estimate of Glass's Δ for group 1
`r(pooled_se_g2)`	pooled SE for Glass's Δ for group 2
`r(pt_est_g2)`	pooled point estimate for Glass's Δ for group 2
`r(ub_g2)`	upper confidence limit for the estimate of Glass's Δ for group 2
`r(lb_g2)`	lower confidence limit for the estimate of Glass's Δ for group 2

All methods:

Macros

`r(by_var)`	grouping variable used in the `by()` statement
`r(varname)`	outcome variable used in the command

2.4 How to use miesize

The `miesize` command calculates two-sample effect sizes from multiply imputed data in wide format. The command can handle situations where either one or both variables (namely, the outcome and grouping variables) are imputed. When neither variable is detected as imputed, then the `esize` command for nonimputed data will be invoked, and a message will be provided to alert the user of this: `It appears that neither of the variables is imputed. The standard 'esize twosample' analysis will be performed. You may wish to check your imputed data are in standard Stata wide format.` The imputed data should be in the wide format that Stata provides. That is, imputed variables are named sequentially as *_m_varname*, where *m* is the imputation number, for example, `_2_price`, where this is the second imputed dataset for the variable `price`. As described in section 1, Rubin's rules are used to estimate the pooled effect size, SEs, and hence the associated confidence intervals around the point estimate. The `miesize` command also operates where only a single imputed value is used for the grouping or outcome variable. `miesize` also handles the situation where either only the outcome variable contains imputed values or only the grouping variable includes imputed values.

The default for `miesize` is to calculate Cohen's *d* and Hedges's *g*. The two methods differ in how they estimate the pooled SD of the two groups, as outlined above. In effect, Cohen's *d* uses the arithmetic mean of the two group variances. In contrast Hedges's *g* uses a weighted average that accounts for the sample sizes of each group. As mentioned in section 1, Glass's Δ is more appropriate where the variance of the outcome varies significantly between the two groups. In Stata, this can be formally tested using the `sdtest` command. In this context, the variance across the groups will be compared. Thus, the form of the command used will be

`sdtest` *varname*`, by(`*groupvar*`)`

where the *varname* is the outcome of interest and the *groupvar* is the group-identifying variable. Where a statistically significant difference between the variances for each group is present, it is likely to be more appropriate to use Glass's Δ to report the effect size, rather than d or g. For Glass's Δ, the two effect sizes (Δ_1 and Δ_2) are given relating to the SDs of the two groups. For experimental data, usually the SD for the control group is used when calculating Δ_1. However, as stated earlier, for observational data, the choice of which of the group's SDs to use appears arbitrary. For dummy-coded group variables (one or zero/absent or present), the group with "zero" status could be the one to have the SD used in calculating Δ_1. However, in the absence of a compelling reason to consistently report a Δ value based on one or the other group's SD, it has been recommended that both effect-size estimates (Δs) be reported for observational data (Kline 2013).

3 Example

For this example, an analysis of data used for evaluating the recruitment and selection process into UK-based general practice postgraduate medical training is used for illustrative purposes (Tiffin et al. 2024).[1] One aim of the analysis was to estimate effect sizes in relation to selection test scores for various demographic characteristics. Clearly, any substantial association with such personal qualities would influence the demographics of the population of doctors in training finally selected. In this context, there were complete data for the first stage of selection, which comprised a situational judgment test (SJT) and clinical knowledge assessment (the clinical problem solving [CPS] test). These scores are combined into a summed total that is used in the selection process. While all candidates had SJT and CPS scores, not all had received a standardized face-to-face selection center (SC) assessment. This was a special case of missing data by design. For some years, those who achieved only low combined scores on the SJT and CPS did not proceed to the SC stage. In addition, for some later years, candidates who had achieved relatively high combined SJT and CPS scores were exempted the face-to-face selection stage. Moreover, the face-to-face selection process was suspended completely during the COVID-19 pandemic. Overall, this means that SC scores were not present for around half the doctors in the sample with first-stage selection assessment scores. Thus, estimating the "true" underlying effect size for each demographic characteristic in relation to the scores for the face-to-face SCs required data imputation. This was performed in Stata using chained equations (Royston and White 2011). The estimates from the imputations stabilized after $m = 5$ imputations, but as a precaution, $m = 10$ imputed datasets were used. For this particular illustrative example, we will estimate the effect size of male gender on SC scores. Using the `sdtest` command in Stata, we see the results indicate that the assumption of equal variance in SC scores across gender groups can be rejected:

1. The dataset used in the example is held within the UK Medical Education Database (UKMED) in a trusted research environment.

```
. sdtest sc, by(male)
Variance ratio test
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	8,689	45.81666	.047401	4.418472	45.72375	45.90958
1	6,260	44.18115	.0605609	4.791592	44.06243	44.29987
Combined	14,949	45.13178	.0380222	4.648829	45.05725	45.20631

```
    ratio = sd(0) / sd(1)                                     f =   0.8503
H0: ratio = 1                           Degrees of freedom = 8688, 6259
    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0000         2*Pr(F > f) = 0.0000           Pr(F > f) = 1.0000
```

Because there is evidence of unequal variances between the groups, we will use Glass's Δ to estimate the effect size using `miesize`:

```
. miesize sc, by(male) glass
Average obs per Group 1 18417
Average obs per Group 2 12800
```

Effect size	Pooled estimate	[95% conf.	interval]
Glass's Delta 1	.41731	.3819939	.4526262
Glass's Delta 2	.3982446	.3644922	.431997

As can be seen from the results, the effect size for male gender is around 0.40 (that is, commonly interpreted as "modest"). As expected, it varies slightly depending on whether the SD of the SC scores for males (0.40) or females (0.42) is used. We can compare our results with those obtained using the `esize twosample` command, which uses only the observed data:

```
. esize twosample sc, by(male) glass
Effect size based on mean comparison
                              Obs per group:
                                   male==0 =      8,689
                                   male==1 =      6,260
```

Effect size	Estimate	[95% conf.	interval]
Glass's Delta 1	.3701538	.3371882	.4030987
Glass's Delta 2	.3413301	.3082789	.3743549

As might be anticipated, the effect sizes for gender are modestly smaller than those estimated from the imputed data. This is because certain candidates that scored either especially high or low on the first stage of selection tests will not have observable SC scores. Given that, in this sample, males, on average, achieve lower scores on the first-stage assessments, this will produce "indirect range restriction". This in turn restricts the range of observable SC scores, especially for females. This effect attenuates the effect size observed, a well-recognized phenomenon in personnel selection studies. Indeed, multiple imputation has been shown to be one way of addressing this (Zimmermann, Klusmann, and Hampe 2017; Mwandigha 2017). Thus, the imputed values will offset

this effect by simulating the unobserved SC scores. Note that, as expected, the confidence intervals for these effect sizes are slightly wider for the results derived by `miesize`, compared with `esize`. This is because, as explained above, the variance between, as well as within, imputed datasets is accounted for. In this case, the former effect is not large. This is because, across the 10 imputed datasets, the imputed SC scores vary little. Indeed, the range of observed SC scores in this sample is 20 to 52. However, between the $m = 10$ imputed datasets, the imputed SC scores only vary by a range of around 2 points. This is because the relationships between the SC scores and the other variables in the sample are relatively strong. For example, the β coefficient for the SC scores regressed on SJT scores in the observed data is 0.30. Thus, the chained equations impute values with a reasonable level of (apparent) certainty. However, as explained earlier, where the variance in imputed values is higher, this effect will be more marked.

4 Conclusions

The d family of effect size estimators calculates standardized mean group differences for continuous variables. The `miesize` command avoids the tedious task of calculating the pooled estimates and SEs for these estimators when working with multiply imputed data. A number of limitations with the current `miesize` command should be acknowledged. At present, neither `miesize` nor `esize` accommodates weights (for example, survey weights). Also, the derivation of SEs for both commands rests on an assumption of normality. One way of addressing this would be the option to derive the SEs via bootstrapping, or other resampling methods, which is not presently included for the commands. Note that the standard Stata command `esize` can be used to calculate effect sizes from unpaired data. This is achieved by calling `esize unpaired`, then by specifying the two variables of interest with a pair of equal signs between them, for example, `esize unpaired mpg1 == mpg2`. This manner of estimating effect sizes in unpaired data does not currently accommodate multiply imputed data in Stata. Thus, these three limitations offer potential areas for future development of these commands.

5 Acknowledgments

Many thanks to Dr. Lewis Paton (the Hull York Medical School) for his feedback on an earlier version of this manuscript. I am also grateful to Drs. Nick Cox (Durham University) and Yulia Marchenko (StataCorp LLC) for feedback on an earlier version of the code for the `miesize` command. I also thank Fraser Wiggins (University of York, Clinical Trials Unit) for advice in relation to the code for the `miesize` command.

6 Programs and supplemental material

To install the software files as they exist at the time of publication of this article, type

```
. net sj 24-3
. net install st0755     (to install program files, if available)
. net get st0755         (to install ancillary files, if available)
```

7 References

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2021. *Introduction to Meta-Analysis*. 2nd ed. Chichester, UK: Wiley.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203771587.

Grissom, R. J., and J. J. Kim. 2001. Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods* 6: 135–146. https://doi.org/10.1037/1082-989x.6.2.135.

Hedges, L. V. 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6: 107–128. https://doi.org/10.3102/10769986006002107.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-analysis*. San Diego: Academic Press.

Heymans, M. W., and I. Eekhout. 2019. *Applied Missing Data Analysis With SPSS and (R) Studio*. Amsterdam: Heymans and Eekhout.

Kline, R. B. 2013. *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*. 2nd ed. Washington, DC: American Psychological Association. https://doi.org/10.1037/14136-000.

Mwandigha, L. M. 2017. Evaluating and extending statistical methods for estimating the construct-level predictive validity of selection tests. PhD thesis, Health Sciences, University of York. https://etheses.whiterose.ac.uk/21267/1/Lazaro_Mwakesi_Mwandigha_PhD_thesis.pdf.

Royston, P., and I. R. White. 2011. Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software* 45: art. 4. https://doi.org/10.18637/jss.v045.i04.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. https://doi.org/10.1002/9780470316696.

Smith, M. L., and G. V. Glass. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist* 32: 752–760. https://doi.org/10.1037/0003-066X.32.9.752.

Sterne, J. A. C., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 338: b2393. https://doi.org/10.1136/bmj.b2393.

Tiffin, P. A., E. Morley, L. W. Paton, and F. Patterson. 2024. New evidence on the validity of the selection methods for recruitment to general practice training: A cohort study. *BJGP Open* BJGPO.2023.0167. https://doi.org/10.3399/BJGPO.2023.0167.

van Ginkel, J. R., M. Linting, R. C. A. Rippe, and A. van der Voort. 2020. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment* 102: 297–308. https://doi.org/10.1080/00223891.2018.1530680.

Zimmermann, S., D. Klusmann, and W. Hampe. 2017. Correcting the predictive validity of a selection test for the effect of indirect range restriction. *BMC Medical Education* 17: art. 246. https://doi.org/10.1186/s12909-017-1070-5.

About the author

Paul A. Tiffin is Professor of Health Services and Workforce Research at the University of York, UK, and a practicing adolescent psychiatrist.