A web-based application for effect sizes and their confidence intervals: standardized mean differences, binary outcomes, variance explained, and two-level multilevel models

James O. Uanhoro[1] & Ann A. O'Connell[1]

[1] The Ohio State University

Author Note

Quantitative Research, Evaluation, and Measurement program, Department of Educational Studies.

Correspondence concerning this article should be addressed to James O. Uanhoro, Room 210, 29 W Woodruff Ave, Columbus, OH 43210. E-mail: uanhoro.1@osu.edu

Abstract

There have been increasing calls for applied researchers to see and utilize effect sizes as the primary outcomes of their research. However, this sometimes places a methodological burden on researchers whose primary interests are substantive. Motivated by a desire to help applied researchers better report effect sizes and their confidence intervals, we developed a web application that allows users to calculate effect sizes and confidence intervals for standardized mean differences, binary outcomes, ANOVA, multiple regression, and multilevel models. In this paper, we demonstrate the use of our application in the context of binary outcomes and multilevel models. It is our hope that through our work, applied researchers can better contribute to a cumulative science.

Keywords: effect size, confidence interval, binary outcomes, multilevel models, web application, calculator

A web-based application for effect sizes and their confidence intervals: standardized mean differences, binary outcomes, variance explained, and two-level multilevel models

Introduction

According to Cohen (1990), the "primary product of a research inquiry is one or more measures of effect size, not p values." Cohen makes this claim because effect sizes provide applied researchers with means to convey the substantive findings from empirical studies. Furthermore, upon considering the utility of effect sizes in facilitating a cumulative science, Wilkinson and the APA Task Force on Statistical Inference (1999) advised that researchers should "always present effect sizes for primary outcomes" (p. 599).

Despite such clarity of guidance, there remained confusion on the meaning of the term. Motivated by a need to clear this confusion, Kelley and Preacher (2012) defined an effect size as "a quantitative reflection of a magnitude of some phenomenon that is used for the purpose of addressing a question of interest." The broadness of this definition is purposeful; it encompasses multiple perspectives on what an effect size is. For one, it goes beyond the standardized mean difference (such as Cohen's d) which is commonly used to facilitate the interpretation of the effect of interventions within the field of education.

Furthermore, given that effect sizes are simply another sample statistic, they are only estimates; thus, it is prudent to communicate the uncertainty about such estimates. One tool that can be used for communication of such uncertainty is the confidence interval. If a research inquiry is worth the allocation of any resources (such as participants' and consumers' time), then it is incumbent on the researcher to report enough information to help consumers of their research appraise the outcome of its inquiry. The Publication Manual of the American Psychological Association (2010) states that "complete reporting for all tested hypothesis and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals" (p. 33).

Given the call for applied researchers to report effect sizes and their confidence intervals, one might assume that the ability to compute a variety of effect sizes alongside

their confidence intervals is readily available in commonplace statistical software. However, SPSS, which continues to be used by a relatively large percentage of applied researchers (Muenchen, 2017), does not return a standardized mean difference when the independent samples t-test is performed. And although it returns an unstandardized mean difference – a credible effect size measure – with confidence intervals, its calculation of the confidence intervals is inadequate if the intervals are to be used to communicate uncertainty about a non-nil effect (Smithson, 2001).

Moreover, although commonplace statistical software return variance-explained effect sizes such as R^2 in regression, and partial eta-squared (η_p^2) in ANOVA, they do not return confidence intervals on these effect sizes. Usually, the diligent applied researcher attempting to compute these confidence intervals has to find, verify, learn to use and modify syntax written for use in SPSS or the researcher's preferred statistical software. This raises the methodological and programming demands on applied researchers whose interests are substantive, rather than methodological.

An alternative strategy researchers employ is to use freely available web calculators that use summary statistics to compute effect sizes. However, it is rare to find online effect size calculators that compute confidence intervals in the manner recommended by Smithson (2001) when they calculate confidence intervals. A rewarding strategy might be to invest in the programming skills required to use statistical computing software like R; however, this also increases the programming demands on applied researchers.

Motivated by a desire to aid applied researchers in their endeavor to meet the aforementioned "minimum expectations" of reporting, we developed a web application to calculate effect sizes for different analyses and study designs.[1] Where feasible, our application returns confidence intervals – using the methods advised by Smithson (2001) – alongside the effect size estimates. Researchers can obtain confidence intervals for the following effect size measures: Cohen's d (single, independent and paired samples), odds

[1] The application is available at https://effect-size-calculator.herokuapp.com/.

ratios, relative risk (reduction), η_p^2, R^2, and the intra-cluster correlation coefficient (ICC).[2] Additional effect size measures the calculator computes are absolute risk (reduction), number needed to treat, and different measures of R^2 in two-level linear multilevel models - see Figure 1 for the range of effect sizes computed by our application.

In the remainder of this text, we examine two groups of effect sizes our application computes: effect sizes for binary outcomes, and R^2 in two-level MLMs. We focus on these two groups of effect sizes as we believe the other effect sizes have received ample coverage in the education literature unlike both these groups.

Binary outcomes

Binary outcomes are often analyzed in the field of epidemiology, where researchers analyze the efficacy of interventions in the treatment of diseases with dichotomous clinical outcomes such as cured vs. diseased. Given this dichotomy, the outcome from a two-group drug trial can be presented using a 2 by 2 contingency table as seen in Table 1. Effect sizes such as the odds ratio (OR), relative risk/risk ratio (RR), absolute risk (AR) and number needed to treat (NNT) can be calculated from such contingency tables.

Greenberg and Abenavoli (2017) discussed how these effect size measures can be used in combination with a conventional metric such as the standardized mean difference to communicate the impact of educational interventions targeted towards the entire population of students. They argue convincingly that when outcomes are dichotomous and/or can be expressed as proportions (% proficient vs. % not proficient), these effect sizes can help us better communicate the effects of our interventions. Measures like the RR and NNT are more intuitive, we will attempt to demonstrate their utility using the worked example below. See Ellis (2010) for an engaging introduction to effect sizes for binary outcomes.

---

[2] We included the ICC even though it is not often considered an effect size. However, information about ICC can be used to facilitate a cumulative science.

Worked example

We adapted the example in Table 1 - a 2 by 2 contingency table - from Spoth et al. (2011). Spoth et al. (2011) reported on the results of an experiment to reduce the prevalence of substance abuse by adolescents - we focus on the abuse for methamphetamine. See Figure 2 for a screenshot of the worked example in our web application. Given the layout in Table 1, we can calculate a number of effect sizes to quantify the effect of the intervention to reduce the abuse of methamphetamine by adolescents. However, we will focus on the relative risk and number needed to treat because they are most intuitive to grasp:

- the relative risk (RR) is the ratio of the probability of abusing methamphetamine for someone in the intervention group to the probability of abusing methamphetamine for someone in the control group:

$$RR = \frac{224/(224 + 5835)}{378/(378 + 5523)} = 0.577 \tag{1}$$

- the number needed to treat (NNT) is the number of individuals that would need to receive the intervention so that one individual does not abuse methamphetamine; it is the inverse of the absolute risk:

$$NNT = \frac{1}{\frac{224}{224+5835} - \frac{378}{378+5523}} = \frac{1}{0.0271} = -36.9 \tag{2}$$

When the outcome is negative, such that a reduction in behaviour is desired, we calculate the *relative risk reduction* (RRR). The RRR is one minus the RR which is 0.423, and we obtain the NNT by multiplying its value by minus one, which gives us 36.9. For the RRR, those who received the intervention were 42.3% (95% CI [32.2%, 50.9%]) less likely to abuse methamphetamine than those who were in the control group - an RRR of 0% would signify a nil effect. For the NNT, 37 adolescents would have to receive the intervention so

that one adolescent does not use methamphetamine - an increase in the NNT corresponds to a weaker effect.

We provided a number of options in our application for computing the point estimates of these effect sizes and their CI. For the computation of confidence intervals around the OR and the RR, the `Unconditional maximum likelihood estimation (Wald)` method - a normal approximation method - suffices for large sample sizes (Jewell, 2004). The `small sample adjustment (small)` method should be used as a diagnostic - if it markedly differs from the Wald method, then the sample size is too small to use the Wald method (Jewell, 2004, p. 85). When this problem occurs for the OR, the `Median-unbiased estimation (mid-p)` method should be used (Agresti, 2013, p. 94). For the RR, `Bootstrap estimation (boot)` may be used as an alternative.

Two-level multilevel models

Educational data is often hierarchical in nature, examples include students within classrooms, and teachers within schools. Researchers often use multilevel models (MLMs) to model these structures. Given the relative familiarity of most applied researchers with R^2 in OLS regression, methodologists have tried to create R^2 measures for MLMs that attempt to recreate the characteristics of OLS R^2. For our application, we built in two sets of R^2 measures: level-one and level-two R^2 (R^2_{SB}) by Snijders and Bosker (1994), and marginal and conditional R^2 (R^2_{NS}) by Nakagawa and Schielzeth (2013). The utility of these R^2 measures is best explained using a demonstrated example.

Demonstrated example

We will use a subsample of the High School and Beyond (HS&B) dataset with 7,185 students nested within 160 schools (Raudenbush & Bryk, 2002) - see Table 2 for summary statistics. We will model math achievement (*mathach*) using minority status (*minority*), and socio-economic status (*SES*) as level 1 predictors of math achievement. We will include school average SES (*meanSES*) and whether the school was Catholic or public (*catholic*) in

our model as level 2 predictors of both the intercept and the slope of *SES*. *SES* will be group-mean centered, while *meanSES* will be grand-mean centered. Our application performs the computations as a random-intercepts model as required for the computation of both R^2_{SB} & R^2_{NS}. See Figure 3 for how to specify this model in our web application.

As seen in Figure 4, the level-one R^2 was 0.206, this implies that we modelled 20.6% of the variability between students; the level-two R^2 was 0.702, this implies that we modelled 70.2% of the variability between schools. The marginal R^2 was 0.206, implying that the fixed effects in our model - *minority*, *SES*, *meanSES*, *catholic*, and the cross-level interactions between *SES* and *meanSES*, and *SES* and *catholic* - explained 20.6% of the variability in math achievement. The conditional R^2 was 0.247, implying that the aforementioned fixed effects together with the random intercept explained 24.7% of the variability in math achievement. In order to calculate R^2_{SB}, our application had to compute an empty model in addition to the model described above - the references to `base model` in Figure 4 are statistics calculated from this empty model.[3]

Additionally, the application creates a text file containing the results of the models (including coefficients and p-values) as seen in Figure 5. The user can save and examine the file to ensure comparability of values to that obtained from their preferred multilevel software. We implore the user to verify that the models converged (see Figure 4). If they did not converge, the user should change the `Optimization Method` using the provided dropdown menu (see Figure 3).

---

[3] The models are fitted using ML to ensure comparability of the base model with the fitted model in the computation of R^2_{SB} (Zuur, Ieno, Walker, Saveliev, & Smith, 2009, p. 122). It would be prudent to use REML to compute R^2_{NS}. However, doing this in our application would require solving three MLMs: a base model and full model using ML for R^2_{SB}, and a full model using REML for R^2_{NS}. Due to computational limitations, we programed our application to solve only two MLMs and calculate R^2_{NS} from the full model fitted using ML.

Conclusion

In the preceding sections, we have outlined some of the capabilities our web application provides. It is our hope that our application helps applied researchers calculate effect sizes so as to better communicate the outcomes of their studies. In the future, we intend to increase the number of effect sizes our application covers. And in the area of MLMs, we hope to further develop the application to allow users specify three-level MLMs, so as to compute related effect sizes.

References

Agresti, A. (2013). Inference for Contingency Tables. In *Categorical data analysis* (pp. 70–114). Wiley-Interscience.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. doi:10.1037/0003-066X.45.12.1304

Ellis, P. D. (2010). Introduction to effect sizes. In *The essential guide to effect sizes : Statistical power, meta-analysis, and the interpretation of research results* (pp. 1–30). Cambridge University Press.

Greenberg, M., & Abenavoli, R. (2017). Universal Interventions: Fully Exploring Their Impacts and Potential to Produce Population-Level Impacts. *Journal of Research on Educational Effectiveness*, *10*(1), 40–67. doi:10.1080/19345747.2016.1246632

Jewell, N. P. (2004). Estimation and Inference for Measures of Association. In *Statistics for epidemiology* (pp. 76–97). Chapman & Hall/CRC.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. doi:10.1037/a0028086

Muenchen, R. A. (2017). The Popularity of Data Science Software. Retrieved from http://r4stats.com/articles/popularity/

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. doi:10.1111/j.2041-210x.2012.00261.x

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical Linear Models* (2nd ed., p. 485). Thousand Oaks: Sage.

Smithson, M. (2001). Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals. *Educational and Psychological Measurement*, *61*(4), 605–632.

doi:10.1177/00131640121971392

Snijders, T. A., & Bosker, R. J. (1994). Modeled Variance in Two-Level Models. *Sociological Methods & Research*, *22*(3), 342–363. doi:10.1177/0049124194022003004

Spoth, R., Redmond, C., Clair, S., Shin, C., Greenberg, M., & Feinberg, M. (2011). Preventing Substance Misuse Through Community–University Partnerships. *American Journal of Preventive Medicine*, *40*(4), 440–447. doi:10.1016/j.amepre.2010.12.012

Wilkinson, L., & the APA Task Force on Statistical Inference, A. P. A. B. of S. A. (1999). Statistical Methods in Psychology Journals. *American Psychologist*, *54*(8), 594–604. doi:10.1037/0003-066X.54.8.594

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer New York. doi:10.1007/978-0-387-87458-6

Table captions

Table 1. Outcome of an RCT to prevent substance mis-use by adolescents after 4.5 years

Table 2. HS&B Descriptive Statistics

Figure captions

Figure 1. Range of effect sizes provided by our application - screenshot from web application

Figure 2. Binary outcome problem with solution - screenshot from web application

Figure 3. Model specification - annotated screenshot of web application

Figure 4. Model results (Effect sizes) - screenshot of web application

Figure 5. Model results (Details) - screenshot of text file produced by web application

Table 1

Outcome of an RCT to prevent substance mis-use by adolescents after 4.5 years

	Abusing Methamphetamine	Not abusing Methamphetamine
Intervention group	224	5,835
Control group	378	5,523

Note. Adapted from Spoth et. al. (2011). Data in table are cell sizes. We calculated the cell sizes using total study sample and the reported prevalence rates. Missing data were handled using Full-information maximum likelihood so we were able to use the total study sample size. This allowed us to calculate confidence intervals; however, these intervals may be overly narrow given our failure to account for the clustered design of the study.

Table 2

HS&B Descriptive Statistics

		mean	sd
Student-level variables (n = 7185)			
Minority status dummy (Minority = 1)	*minority*	0.27	0.45
Socio-economic status	*ses*	0.00	0.78
Math achievement	*mathach*	12.75	6.88
School-level variables (n = 160)			
School average SES	*meanses*	0.00	0.41
Catholic or public school dummy (Catholic = 1)	*catholic*	0.49	0.50

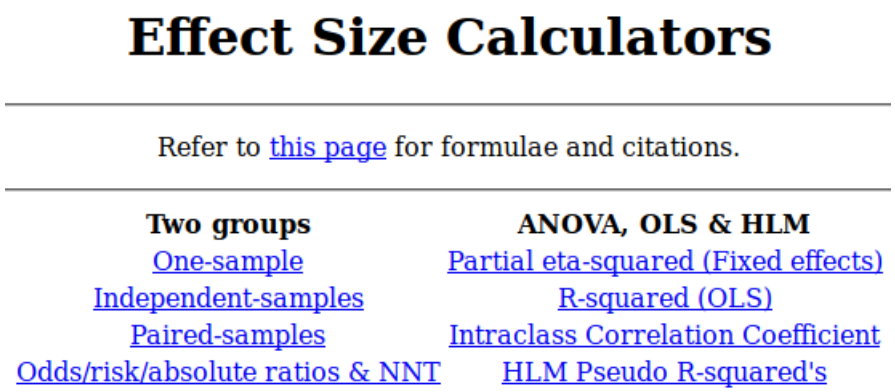

Figure 1. Range of effect sizes provided by our application - screenshot from web application

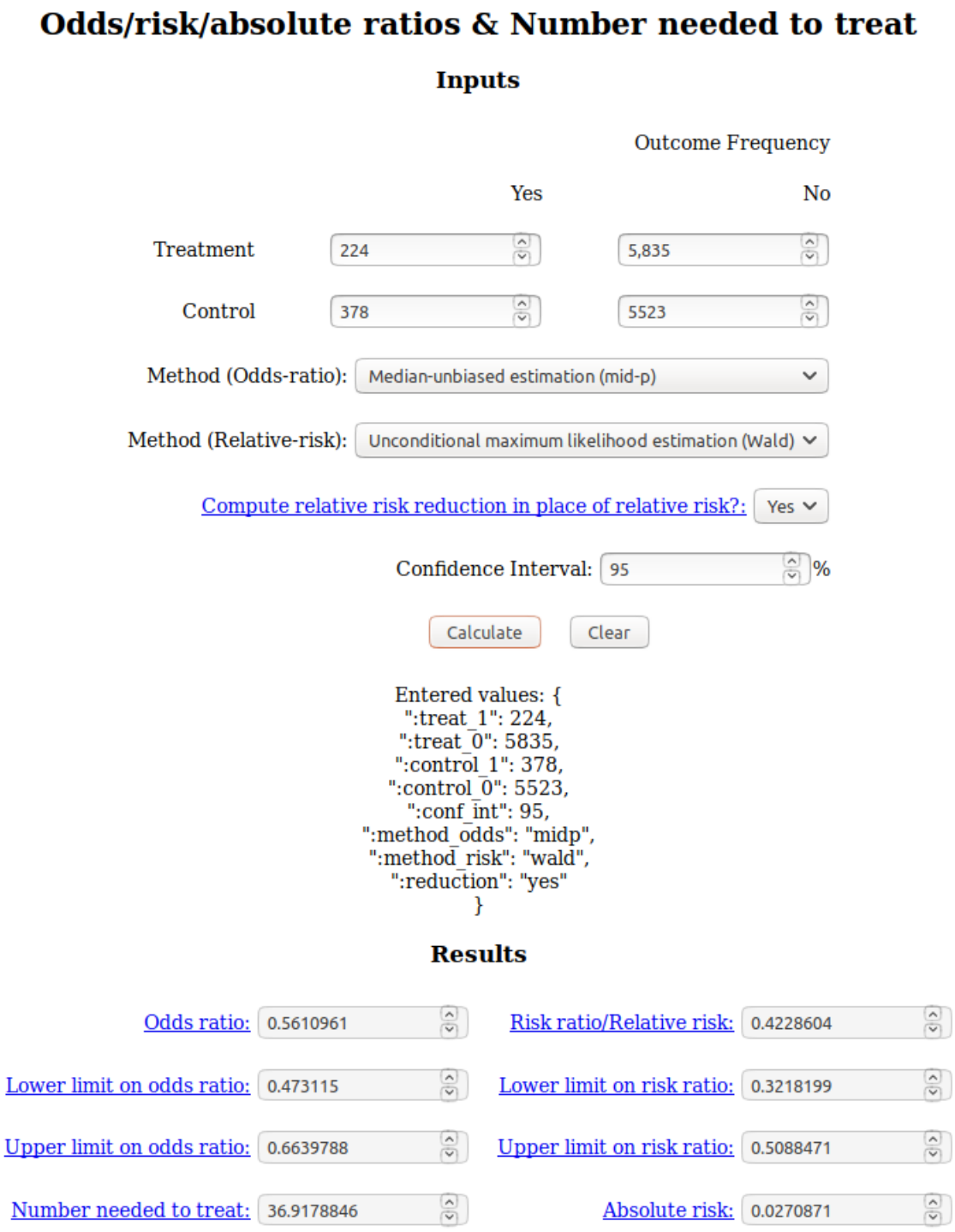

Figure 2. Binary outcome problem with solution - screenshot from web application

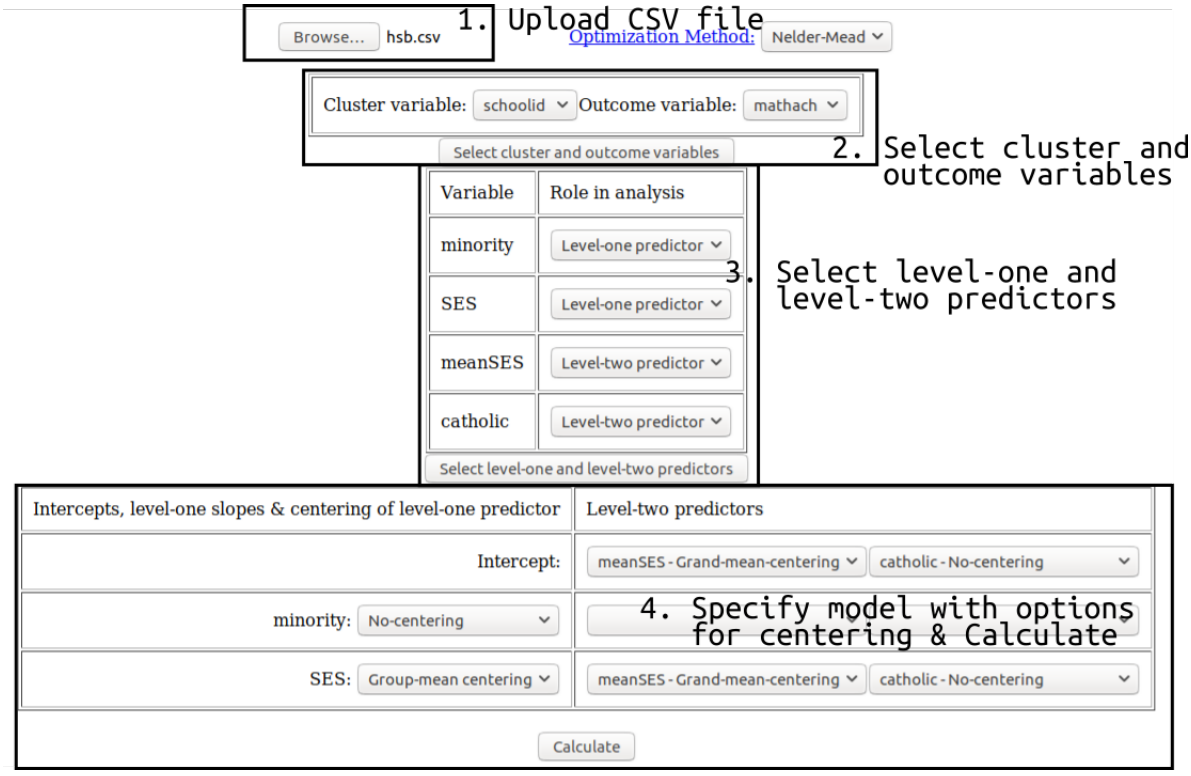

Figure 3. Model specification - annotated screenshot of web application

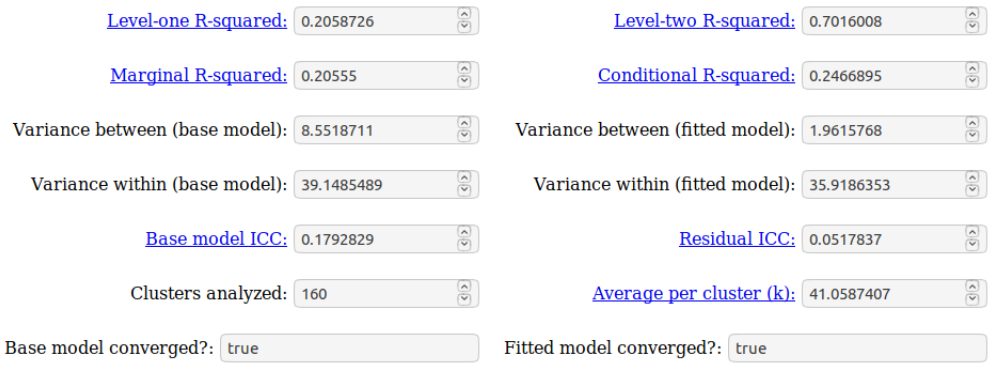

Figure 4. Model results (Effect sizes) - screenshot of web application

```
Base model:
          Mixed Linear Model Regression Results
=========================================================
Model:            MixedLM Dependent Variable: mathach
No. Observations: 7185    Method:             ML
No. Groups:       160     Scale:              39.1485
Min. group size:  14      Likelihood:         -23557.9051
Max. group size:  67      Converged:          Yes
Mean group size:  44.9
---------------------------------------------------------
           Coef.   Std.Err.    z     P>|z|  [0.025  0.975]
---------------------------------------------------------
Intercept  12.637     0.244  51.876  0.000  12.160  13.115
groups RE   8.552     0.173
=========================================================

Fitted model:
                        Mixed Linear Model Regression Results
================================================================================================
Model:                        MixedLM          Dependent Variable:          mathach
No. Observations:             7185             Method:                      ML
No. Groups:                   160              Scale:                       35.9186
Min. group size:              14               Likelihood:                  -23158.3362
Max. group size:              67               Converged:                   Yes
Mean group size:              44.9
------------------------------------------------------------------------------------------------
                                           Coef.  Std.Err.    z    P>|z| [0.025 0.975]
------------------------------------------------------------------------------------------------
Intercept                                  12.675    0.190  66.715 0.000 12.302 13.047
minority                                   -2.748    0.203 -13.553 0.000 -3.145 -2.350
SES_efc_centered_1                          2.641    0.151  17.481 0.000  2.345  2.938
meanSES_efc_centered_2                      4.181    0.354  11.825 0.000  3.488  4.874
catholic                                    1.675    0.287   5.843 0.000  1.113  2.236
SES_efc_centered_1:meanSES_efc_centered_2   0.998    0.288   3.469 0.001  0.434  1.562
SES_efc_centered_1:catholic                -1.476    0.231  -6.396 0.000 -1.928 -1.024
groups RE                                   1.962    0.054
================================================================================================

A note about modified variable names
_efc_centered_1 after a variable name signifies group-mean centering;
_efc_centered_2 after a variable name signifies grand-mean centering.
```

Figure 5. Model results (Details) - screenshot of text file produced by web application