Tutorials in Quantitative Methods for Psychology
2009, Vol. 5(1), p. 25-34.

DOI: 10.20982/tqmp.05.1.p025

Computing Effect Size Measures with ViSta - The Visual Statistics System

Rubén Daniel Ledesma,
CONICET / Universidad Nacional de Mar del Plata

Guillermo Macbeth
CONICET / Universidad del Salvador, Argentina

Nuria Cortada de Kohan
Universidad de Buenos Aires, Argentina

Effect size measures are recognized as a necessary complement to statistical hypothesis testing because they provide important information that such tests alone cannot offer. In this paper we: a) briefly review the importance of effect size measures, b) describe some calculation algorithms for the case of the difference between two means, and c) provide a new and easy-to-use computer program to perform these calculations within ViSta "The Visual Statistics System". A worked example is also provided to illustrate some practical issues concerning the interpretation and limits of effect size computation. The audience for this paper includes novice researchers as well as ViSta's user interested on applying effect size measures.

In psychological research, Effect Size (ES) measures constitute a necessary complement to statistical significance hypothesis testing (Thompson, 1994, 1998). In this work we: (a) review the importance of ES measures; (b) describe some calculation algorithms used to estimate these measures in case of a difference between two means; and (c) present an easy-to-use computer software to perform these calculations within the *ViSta* statistical system. It is hoped this paper will help increase awareness of these methodologies and facilitate access to the IT tools necessary for their application.

Effect Size Measures

In psychological research, ES represents a way to measure or quantify the effectiveness of an intervention, treatment or program. ES can also be described as the degree of falsity of the null hypothesis (Descôteaux, 2007). This quantification is required for determining sample sizes and to achieve correct statistical decisions (Wilson Van Voorhis & Morgan, 2007). To illustrate the importance of ES, we will analyze an example taken from Moore & McCabe (1993), which is available as a data archive in the *ViSta* examples folder.

Suppose we wish to study the effect of a new teaching activity on the reading skills of students. A study using two groups is undertaken. The new teaching activity is applied with the subjects of one group (the experimental group), while the conventional teaching activity is applied with the subjects of the other group (the control group). Afterwards, both groups are given a reading test, with the scores of the reading test constituting the dependent variable Y in the study. Table 1 shows the results of the experiment.

In this case, the results of the *t*-test shows a significant difference between the means of the groups, leading the researcher to reject the null hypothesis that predicted equal means, or the "0" effect of the treatment (the new teaching activity). But, what is the magnitude of the observed difference? Is this difference significant in practical terms? To what extent is the new teaching activity better? These

Rubén Daniel Ledesma, Río Negro 3922, Mar del Plata (7600), Argentina, rdledesma@gmail.com, tel:+ 54 223 4752266. A Spanish tutorial for a preliminary version of this software has been published in Ledesma, Macbeth & Cortada de Kohan (2008).

Table 1. Result from the hypothetical experiment from Moore & McCabe (1993).

Group	*n*	*Mean*	*S.D.*
Experimental	21	51.476	11.007
Control	22	39.545	14.628

$t(41) = 3.01$, $p < .01$

types of questions can be answered applying ES measures.

It is worth noting that statistical significance does not necessarily inform the researcher about the importance or magnitude of the effect. The classical hypothesis testing model seeks to determine whether or not to reject the hypothesis that maintains that the effect is non-existent (Frías-Navarro, Llobell & García-Pérez, 2000; Gigerenzer, 1993). Therefore, if the null hypothesis is rejected, the researcher can only conclude that the effect is significantly different from "0", which, for all practical matters, is of limited usefulness (Krueger, 2001). Furthermore, statistical significance is not a direct indicator of ES, but rather a functional relation between the sample size, the ES and the *p* value (Descôteaux, 2007). For this reason, a weak ES may appear as statistically significant if the sample size is sufficiently large; and, conversely, an effective intervention may not appear as statistically significant if the sample size is small (Wilson Van Voorhis & Morgan, 2007).

A better indicator of the impact of the new teaching activity can be obtained through a standardized measure of the difference between the means of the groups. For example, the following measure could be applied (Cohen, 1969, 1988, 1994):

$$d = \frac{\bar{Y}_e - \bar{Y}_c}{\sigma} \quad (1)$$

In this equation, $\bar{Y}_e$ and $\bar{Y}_c$ represent the means of the dependent variable Y of the experimental and control groups, respectively, and σ is the average standard deviation for both groups, that is, $\sqrt{11.007^2 + 14.628^2 / 2} = 12.945$. In accordance with the example illustrated in Table 1, we obtain:

$$d = \frac{51.476 - 39.545}{12.945} = 0.922$$

This standardized measure of the difference between the means known as Cohen´s *d* constitutes a possible estimation of the ES, and offers various practical advantages. First, it is easier to work with, since it can be interpreted simply as a *z* score. It indicates the difference between the groups in units of standard deviation. For example, if $d = 1$, this means that the mean of the experimental group is 1 standard deviation away from the mean of the control group. If we consider *d* as a *z* score, we can also apply the transformation to percentiles and obtain an alternative interpretation. Continuing with the same example, we can state that the distribution of the experimental group's scores betters the distribution of the control group's scores by 82%, because that is the area under the normal curve that corresponds to a *z* score = .92. Another important advantage of this ES measure is that it provides a common measuring stick to compare the relative importance of interventions and programs across different research studies, e.g. in meta-analytical studies (Anderson, 1999).

Some Limitations on the Use of Effect Size

Despite the advantages of ES measures, many authors have noted that their use is limited in practice (Coe, 2002; Descôteaux, 2007; Frías- Navarro et al., 2000; Alhija & Levy, 2008). This is so even though some institutions, like the American Psychological Association, have recommended and promoted their use (Thompson, 1998). Similarly, many publications presently require researchers to provide ES measures together with their statistical significance tests (Hunter & Schmidt, 2004). A recent review on ES reporting practices in 10 educational research journals in the years 2003 and 2004 found no difference between journals that require ES reports and journals that have no such policy (Alhija & Levy, 2008). Although the ES estimates were similarly reported in both, the discrepancies between p-value drawn conclusions and ES drawn conclusions were not often discussed. Sun (2008) conducted a compressive review on ES reporting practices of 1,243 studies published in 14 academic journals from 2005 to 2007 and found that 49.1% of the articles reported ES and 56.7% of them interpreted ES. The author concludes that "it is necessary for the academic journals, leading scholars, and academic associations to continue to urge the improvement of effect size reporting and interpreting practices". In the real world there are likely various explanations for why ES measures are not commonly used. Historical circumstances (Descôteaux, 2007), the lack of ES in the most popular statistical software packages and the absence of the topic in courses and manuals (Coe, 2002) explain, in part, the infrequent use of these methodologies.

Effect Size Measures: The Difference-Between-Two-Means Case

To analyze the magnitude of the effect in our example we could simply compare the mean of the dependent variable Y in the experimental group to its counterpart in the control group in order to determine if there is a difference (*di*) between them (Equation 2):

$$di = \bar{Y}_e - \bar{Y}_c \quad (2)$$

The difference (*di*) between the means of both groups generated by equation 2 is not stable and homogenous because it depends on the unit of measure of the dependent variable. This raw difference (*di*) is much too unreliable to provide any useful information, and so it behooves the researcher to standardize it in some way. As we shall see below, there are various possible ways to achieve this objective.

The Most Common Approaches for Estimating ES

Glass's Delta

The difference *di* becomes more useful if it is changed into a *z* score when it is standardized. One possible approach to standardize the difference is shown in equation 3, where the difference between the means is divided by the standard deviation of the control group (S_c):

$$Delta = \frac{\overline{Y}_e - \overline{Y}_c}{S_c} \quad (3)$$

This formula, known as *Glass's Delta* (Glass, McGaw & Smith, 1981), can be used as an estimator of the population parameter Δ of equation 4:

$$\Delta = \frac{\mu_e - \mu_c}{\sigma_c} \quad (4)$$

In equation 4, the values μ_e and μ_c refer to the population means of the dependent variable Y in the experimental and control groups, respectively. σ_c refers to the population standard deviation of the control group. Δ is the population parameter that is being estimated through the calculation of the sample statistic in equation 3.

Hedges's g

Glass's Delta standardizes the difference between the groups through the standard deviation of the control group S_c, as indicated in Equation 3. Nonetheless, the gross difference between the means depends on the variance of both groups. For this reason, *Glass's Delta* is only slightly affected by differences in variability between the experimental and control groups. This characteristic can generate bias in the ES estimation when the variability within each group is different. This is why Hedges proposed changing the standard deviation of the experimental group S_c for a measure based on the variability of both groups (Grissom & Kim, 2005). This path provides a pooled standard deviation S_p by combining the data from the experimental and control groups in a single measure that does not assume variance homogeneity.

The pooled standard deviation in the Hedges' S_p formula is calculated via equation 5:

$$S_p = \sqrt{\frac{(n_e - 1)S_e^2 + (n_c - 1)S_c^2}{n_e + n_c - 2}} \quad (5)$$

S_p accounts for both the internal variability of each group (S_e^2, S_c^2) as well as the size of each group (n_e, n_c) when estimating ES. This measure is less biased than Glass's Delta when not assuming equal variances. The use of the pooled standard deviation S_p to calculate ES when comparing two independent groups is known as Hedges' g (Equation 6):

$$g = \frac{\overline{Y}_e - \overline{Y}_c}{S_p} \quad (6)$$

Hedges' g is an estimation of the corresponding population G, indicated in Equation 7:

$$G = \frac{\mu_e - \mu_c}{\sigma} \quad (7)$$

Both Glass's Delta and Hedges' g have a positive bias, which means they overestimate the ES. To adjust for this bias, Hedges proposed a g_{ajust}, which is calculated using Equation 8.

$$g_{ajust} = g\left[1 - \frac{3}{4df - 1}\right] \quad (8)$$

The greater the degrees of freedom *df*, the lesser the adjustment necessary to estimate a less biased ES, as can be deduced from Equation 8.

Cohen's d

Cohen's *d* (1988, 1994) is one of the most widely used measures in specialized publications to calculate ES, and in meta-analytical studies (Anderson, 1999; Hunter & Schmidt, 2004). To calculate it, see Equation 1. Cohen's *d* can also be calculated from *t*-test results (Thalheimer & Cook, 2002). For example, knowing the *t* value and the size of each group, the equation would be:

$$d = t\sqrt{\left(\frac{n_e + n_c}{n_e n_c}\right)\left(\frac{n_e + n_c}{n_e + n_c - 2}\right)} \quad (9)$$

This type of conversion is useful to compute ES from research papers that only report results based on *t*-test.

The alternative uses of Cohen's *d*, Hedges' *g* or Glass's Delta depend on the properties of the standard deviation of the two compared groups. It is assumed that both standard deviations are estimates of the same population value when *d* and *g* are calculated (Coe, 2002). When the difference between both does not depend only on sampling variation, then the standard deviation of the control group and the calculation of Glass's Delta would be a better choice. In this case, the variability of the group that was not affected by any experimental manipulation gives a more accurate approximation to the population standard deviation.

Other ES Measures

The CLES Statistic

McGraw and Wong (1992) propose another method to estimate the ES when comparing two means of independent

samples: the *CLES* (Common Language Effect Size) statistic. This statistic is easier to interpret than the others, given that the magnitude of the difference is expressed as a probability. More precisely, the CLES statistic estimates the probability that a randomly selected individual from the experimental group will have a higher score than a randomly selected individual from the control group (Valera-Espín & Sánchez-Meca, 1997). To calculate it, the *z* score from Equation 10 is needed:

$$Z = \frac{\left|\overline{Y}_e - \overline{Y}_c\right|}{\sqrt{S_e^2 + S_c^2}} \qquad (10)$$

Afterwards, it has to be found in the typical normal distribution, the probability of a value less than the one obtained in the previous equation. In the proposed example, this would be:

$$Z = \frac{51.476 - 39.545}{\sqrt{11.007^2 + 14.628^2}} = 0.652 \text{, and } p(Z < 0.652) = 0.743$$

This result is easily interpreted, i.e. 74.3% of the time, a randomly picked subject from the experimental group will have a value greater than a randomly picked subject from the control group. Further, this conversion of the ES to a probability could also be applied to other standard forms of ES estimation, such as Cohen's *d*, so as to have a more universal form of interpretation.

d-to-r Conversion

Another measure that is direct and simple to interpret is the conversion of Cohen's *d* to *r*. The latter is the biserial correlation between an independent binary variable *X* and a dependent numeric variable *Y* (Cohen, 1988). *X* has two possible values (for example, 1 and 0), depending on whether it is associated with a participant from the experimental group (*X* = 1) or the control group (*X* = 0). The estimation of ES through the use of *r* has some advantages over the previously mentioned estimations; most notably, it is much easier to interpret. One important advantage of *r* over *d* is the bounded condition of the former. Cohen's *d* behaves like a *z* score but *r* moves between -1 and +1. This property facilitates the interpretation of *r* estimates of ES.

Cohen (1988) proposes Equation 11 to convert *d* to *r*.

$$r = \frac{d}{\sqrt{d^2 + (1/pq)}} \qquad (11)$$

The *p* and *q* values correspond to the proportion of subjects belonging to the experimental and control groups, respectively. In our example, the standardized difference between means is *d* = 0.922. Inputting the corresponding values in Equation 11, we have:

$$r = \frac{0.922}{\sqrt{0.922^2 + (1/0.488 \times 0.512)}}$$
$$= \frac{0.922}{\sqrt{0.850 + (1/0.249)}} = \frac{0.922}{2.20} = 0.42$$

It can be observed that the greater the *d* value, the greater the biserial correlation between *X* and *Y*. Also, the greater the discrepancy between *p* and *q* – which is to say, between the sizes of the experimental and control groups– the greater the value in the denominator in Equation 11, and so the lesser the *r* correlation.

When the size of the groups is identical ($n_e = n_c$), the value of the term (*1 / pq*) is 4 (1 / (0.5 x 0.5) = 1/ 0.25 = 4). For this reason, when the experimental and control groups are the same size, Equation 11 can be simplified as:

$$r = \frac{d}{\sqrt{d^2 + 4}} \qquad (12)$$

In the given example, the difference in size between the groups is very small, and so Equation 12 yields the same *r* value:

$$r = \frac{0.922}{\sqrt{0.922^2 + 4}} = \frac{0.922}{\sqrt{4.85}} = \frac{0.922}{2.202} = 0.42$$

Rosenthal & Rubin (1982) suggest an alternative form to present and interpret the *d*-to-*r* conversion, which they call the *binomial effect size display* (BESD). With this method, if the outcome variable is also reduced to a dichotomous variable, *r* can be interpreted simply as a difference between proportions (Randolph & Edmondson, 2005).

A Non-Parametric Method: Cliff's Delta

In all the above-mentioned cases, the ES measure is sensitive to violations of the assumption of normality. A more robust measure for these cases has been proposed by Cliff (1993). His approach is different, given that neither means nor standard deviations are used in the calculation; instead, what is considered is essentially the ordinal rather than the interval properties of the data (Hess & Kromrey, 2004). Specifically, the Cliff's Delta statistic is expressed as:

$$Cliff's \;\; Delta = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{n_1 n_2} \qquad (13)$$

Where x_1 and x_2 are scores within group 1 and group 2, and n_1 and n_2 are the sizes of the sample groups. This statistic estimates the probability that a value selected from one of the groups is greater than a value selected from the other group, minus the reverse probability. Cliff understands that this is a measure of dominance, a concept that refers to the degree of overlap between two distributions. An effect size of 1.0 or -1.0 indicates the absence of overlap between the two groups, whereas a 0.0 indicates the group distributions overlap completely.

This measure can be used when the data distribution deviate greatly from the normal model, or when the variable being compared corresponds to an ordinal level of measurement. The non-parametric nature of Cliff's Delta reduces the influence of factors such as the groups' variance

Table 2. Interpretations of effect sizes (taken from: Coe, 2002).

Effect Size	Percentage of control group who would be below average person in experimental group	Rank of person in a control group of 25 who would be equivalent to the average person in experimental group	Probability that you could guess which group a person was in from knowledge of their 'score'	BESD	CLES
0.0	50%	13th	0.50	0.00	0.50
0.1	54%	12th	0.52	0.05	0.53
0.2	58%	11th	0.54	0.10	0.56
0.3	62%	10th	0.56	0.15	0.58
0.4	66%	9th	0.58	0.20	0.61
0.5	69%	8th	0.60	0.24	0.64
0.6	73%	7th	0.62	0.29	0.66
0.7	76%	6th	0.64	0.33	0.69
0.8	79%	6th	0.66	0.37	0.71
0.9	82%	5th	0.67	0.41	0.74
1.0	84%	4th	0.69	0.45	0.76
1.2	88%	3rd	0.73	0.51	0.80
1.4	92%	2nd	0.76	0.57	0.84
1.6	95%	1st	0.79	0.62	0.87
1.8	96%	1st	0.82	0.67	0.90
2.0	98%	1st (or 1st out of 44)	0.84	0.71	0.92
2.5	99%	1st (or 1st out of 160)	0.89	0.78	0.96
3.0	99.9%	1st (or 1st out of 740)	0.93	0.83	0.98

differences or the presence of outliers.

Interpreting Effect Size

This section summarizes possible ES interpretations according to Coe (2002). Table 2 shows Coe's data (reproduced here with the author's permission). Column 1 lists possible ES values from *d*, Delta or *g* (they can be interpreted like *z* scores). Next follow the percentile conversion (column 2) and the equivalent change in rank order for a group of 25 (column 3). For example, for an effect size of 0.9 (approximately that of the example at the beginning of this paper), the value of 82% indicates that the average person in the experimental group would score higher than 82% of the control group. If the control group consisted of 25 participants, this would be the same as saying that the person ranked 5th in this group would be equivalent to the average person in the experimental group.

The fourth column of Table 2 shows another way of describing the overlap between the two groups. It refers to the probability that one could guess which group a person came from based solely on their test score. This probability equals 0.50 if both groups overlap completely, which means ES equals zero. The probability of guessing correctly increases as the ES increases. In our example, with a difference between the two groups equivalent to an effect size close to 0.90, the probability would be 0.67.

As previously indicated, if the dependent variable is reduced to a variable with two categories, the BESD method can be used to interpret ES as a difference in the proportions in each category. In our example, this value is 0.41, which means 20% of the control group and 61% of the treatment group reached some threshold of success. Lastly, column 6 shows the CLES statistic, which is interpreted as previously mentioned.

Besides these statistical criteria, some practical rules for interpreting ES have been suggested. For example, Cohen (1988) describes an ES value of approximately 0.2 as "small"; an ES value of 0.5 as "medium" and "large enough to be visible to the naked eye"; and an ES value of 0.8 as "grossly perceptible and therefore large". Nevertheless, the value of this rule to applied research has been questioned (Glass et al., 1981), since the practical importance of the effect size

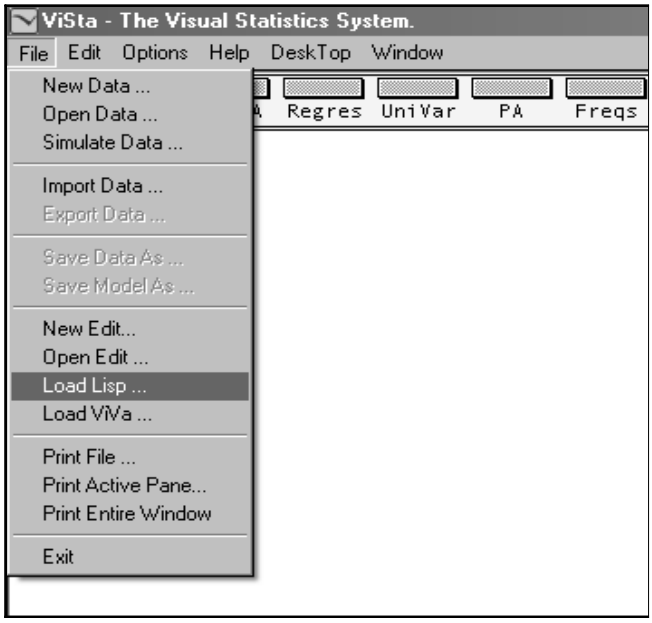

Figure 1. Load the *ES-calc* file into *ViSta* using the "*Load Lisp*" command.

also depends on other variables, such as the effectiveness of other, alternative treatments and the cost-benefit analysis of the treatment.

To summarize, ES cannot be interpreted the same way in all cases. A single effect size measure can have different practical meanings depending on the specific problem being evaluated. For this reason, in each case, relevant theoretical and practical aspects should be considered for the problem being studied. In addition, when the ES estimates and the p-value interpretations lead to different conclusions, assumptions about the frequency distributions and standard deviation properties shuld be carefully revised (Alhija & Levy, 2008).

Calculating ES Measures in Vista

Description of the Es-calc Module

Es-calc is a module that can be integrated into the *ViSta* (Young, 1996) environment, and that can be used to calculate ES measures from raw data or with a calculator by inputting the means, standard deviations and sample sizes. In both instances, the *ViSta* statistical system is required. *ViSta* is a free, expandable statistics program that can be used as a platform for the development of new methods or to expand the system's pre-existing methods. *ViSta* was created by Professor Forrest W. Young at the L. L. Thurstone Psychometric Laboratory (University of North Carolina, Chapel Hill). It is an open-source project on which several developers collaborate.

At a more technical level, we have utilized *XlispStat* (Tierney, 1990) to develop the program's calculation functions and graphic user interface. *XlispStat* is the programming language underlying the *ViSta* system. Readers interested in a general review of the capabilities and functionality of *ViSta* may consult Molina-Ibañez, Ledesma, Valero-Mora & Young (2005). A more detailed review of the program may be found in Young, Valero-Mora & Friendly (2006).

Application Screenshots

Figure 1 shows how to upload the Effect size functions into the ViSta environment (See Appendix for more details on how to download and install ViSta and ES-Calc).

DataSheet Editor - Ed/Cd/Reading.dsh#1

Cd/Reading.cls	DRP-Score	Group
Class[43 X 2]	Numeric	Category
Group[Treatment]	24.00	Treatment
Group[Treatment]	43.00	Treatment
Group[Treatment]	58.00	Treatment
Group[Treatment]	71.00	Treatment
Group[Treatment]	43.00	Treatment
Group[Treatment]	49.00	Treatment
Group[Treatment]	61.00	Treatment
Group[Treatment]	44.00	Treatment
Group[Treatment]	67.00	Treatment
Group[Treatment]	49.00	Treatment
Group[Treatment]	53.00	Treatment
Group[Treatment]	56.00	Treatment
Group[Treatment]	59.00	Treatment
Group[Treatment]	52.00	Treatment
Group[Treatment]	62.00	Treatment
Group[Treatment]	54.00	Treatment
Group[Treatment]	57.00	Treatment
Group[Treatment]	33.00	Treatment
Group[Treatment]	46.00	Treatment
Group[Treatment]	43.00	Treatment
Group[Treatment]	57.00	Treatment
Group[Control]	42.00	Control
Group[Control]	43.00	Control
Group[Control]	55.00	Control
Group[Control]	26.00	Control
Group[Control]	62.00	Control
Group[Control]	37.00	Control
Group[Control]	33.00	Control
Group[Control]	41.00	Control
Group[Control]	19.00	Control
Group[Control]	54.00	Control
Group[Control]	20.00	Control
Group[Control]	46.00	Control

Figure 2. Partial screenshot of *ViSta* data file.

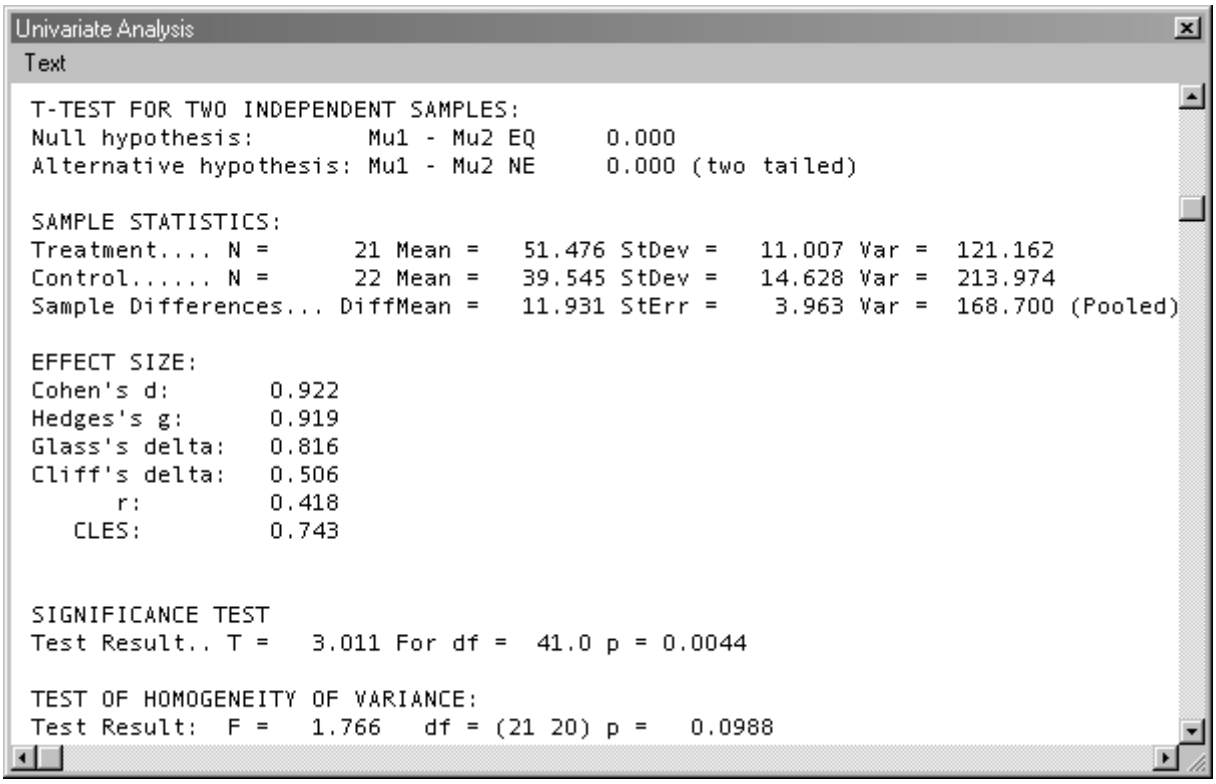

Figure 3. Partial screenshot of statistical report generated in *ViSta*

Figure 2 shows a partial screenshot of ViSta with data corresponding to the above-mentioned example. This type of data file can be created in *ViSta* using the data editor or, also, by importing the data in text format.

In *ViSta*, an ES estimate is performed automatically when a test for the difference between the means of two independent samples is applied. By its nature, this analysis only accepts data entered with a binary independent variable—the two groups being compared—and a quantitative dependent variable, as the data in the example show. After running the command, ViSta provides a report of results, as shown in Figure 3. This figure shows the report with the basic statistical results for the mean comparison for the example's data. The first part includes descriptive information (group sizes, means, standard deviations, etc.), while the second part shows different ES estimates. Lastly, *t*-test and homogeneity of variance test results are displayed.

Calculating ES from Means, SDs and Sample Sizes

The example we have been using for demonstrative purposes has a complete data set, but in some cases, the researcher may not have this information. This would be expected, for instance, in meta-analytical research. For such instances, the *ES-calc* module makes it possible to perform the analysis by using an option that only requieres summary data. This is available from the *ES-calc* item that appears on the *ViSta's* main menu (Figure 4). By selecting *"Effect Size from Means and SDs"*, a dialog box will open, prompting the user to enter the data required to calculate ES measures (see Figure 5). After providing the data and clicking OK, a report of the results will appear (see Figure 6). As can be seen in Figure 4, the program also has the ability to calculate Cohen's *d* from the *t*-test value. For this purpose, the Equation 9 conversion is used.

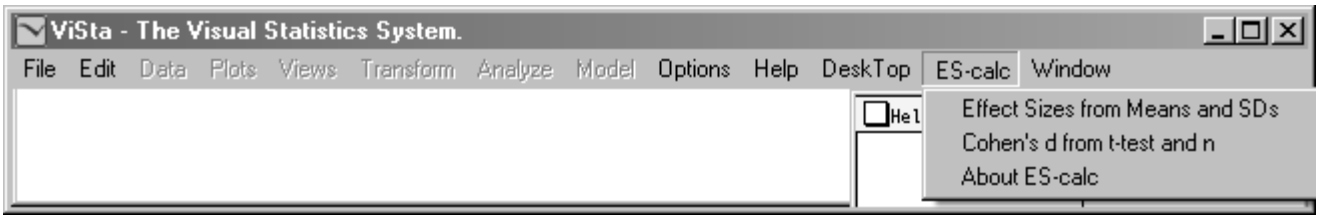

Figure 4. Finding *ES-calc* in *ViSta's* main menu.

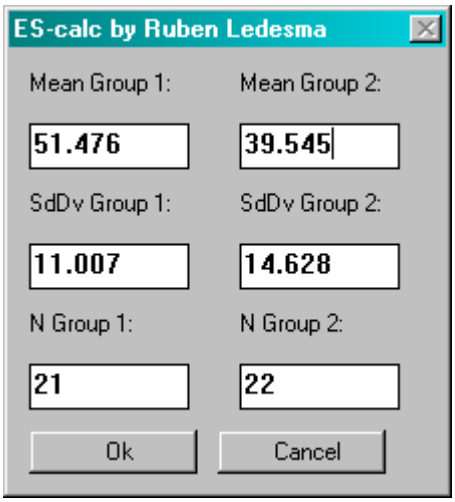

Figure 5. *ES-Calc* dialog box.

programs. Similarly, the most popular statistics programs do not always have analysis options for ES measures clearly displayed in their menus.

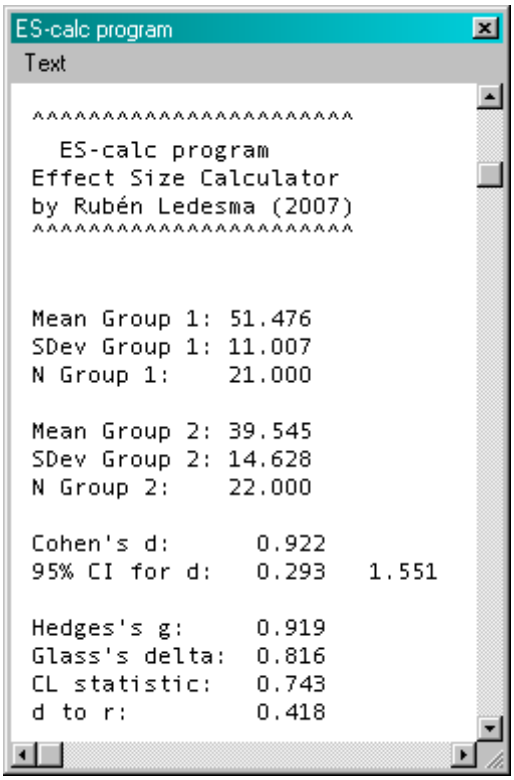

Figure 6. *ES-Calc* report of results.

Conclusion

Presently, there is general agreement in highlighting the importance of ES as a necessary complement to hypothesis testing methods and for determining sample sizes (Descôteaux, 2007; Wilson Van Voorhis & Morgan, 2007). For this reason, experts and the editorial requirements of specialized journals are strongly encouraging the use of these techniques. ES measures allow for a more direct appreciation of the magnitude of the phenomena being studied, and provide a way to interpret results more clearly. In the field of psychological research, including ES in the analysis of data allows for more informed decision-making and more appropriate evidence-based decisions. Additionally, these measures are a key and necessary factor for the integration of results in meta-analytical studies (Hunter & Schmidt, 2004).

Despite the value of these methods, it is clear that the use of ES is not widely practiced in educational and psychological research (Alhija & Levy, 2008; Sun, 2008; Dunleavy, Barr, Glenn & Miller, 2006). This can be attributed to a lack of awareness about these techniques, among other reasons. Many frequently used applied statistics manuals do not include them in their main content, nor are they often included in graduate and post-graduate educational

In this context, this paper aims to contribute to the efforts of institutions such as the APA to raise awareness and encourage the use of ES in psychological research. For this purpose, we have herewith provided a review for the case of the difference between two means, and presented a software application for the *ViSta* statistics program that is free and easy to use. We hope that this tool may help complement the application of hypothesis testing mehods and in this way promote the inclusion of ES measures in empirical studies. Additionally, thanks to its simplicity, we believe that *ES-calc* can also be useful as an educational tool for teaching statistics. Planned expansions of Es-Calc include: 1) further ViSta developments towards ES indexes for categorical data, 2) confidence intervals for ES measures, and 3) statistical data analysis tools for meta-analytical applications.

Lastly, we wish to warn about certain problems that could affect the use and interpretation of ES measures in practice (Coe, 2002). With the exception of Cliff's Delta, the other measures are based on the assumption of normal distributions and equal variances in the groups. Furthermore, the results could be affected when repeated measures are used (Algina & Keselman, 2003), the sample has restricted range, the distribution is skewed, or outliers

are present in the data set. For these reasons, we recommend that the user examine the data for these types of problems before applying parametric ES measures. ViSta provides many alternative graphic resources to do this, including dynamic histograms, normal probability plots, box plots, etc.

References

Algina, J. & Keselman, H.J. (2003). Approximate Confidence Intervals for Effect Sizes. Educational and Psychological Measurement, 63, 4, 537-553.

Alhija, F.N. & Levy A. (2008). Effect Size Reporting Practices in Published Articles. Educational and Psychological Measurement (in press).

Anderson, G. (1999). The Role of Meta-Analysis in the Significance Test Controversy. *European Psychologist, 4(2),* 75-82.

Cliff, N. (1993). Dominance statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin, 114,* 494-509.

Coe, R. (2002). It's the Effect Size, Stupid. What Effect Size is and Why it is Important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002.

Cohen (1969). *Statistical Power Analysis for the Behavioral Sciences.* San Diego, CA: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences. Second Edition.* Hillsdate, NJ: LEA.

Cohen, J. (1994). The Earth Is Round ($p<.05$). *American Psychologist, 49,* 997-1003.

Cousineau, D. (2007). Computing the Power of a t Test. *Tutorials in Quantitative Methods for Psychology, 3,* 2, 60-62.

Descôteaux, J. (2007). Statistical Power: An Historical Introduction. *Tutorials in Quantitative Methods for Psychology, 3,* 2, 28-34.

Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. M. (2006). Effect size reporting in applied psychology: How are we doing?. *The Industrial- Organizational Psychologist, 43,* 4, 29-37

Frías-Navarro, M. D. Llobell, J.P & García-Pérez. J.F (2000) Tamaño del Efecto del Tratamiento y Significación Estadística. *Psicothema,12,* 236-240.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in Statistical Reasoning. En G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 311-339). Hillsdale, NJ: LEA.

Glass, G.V., McGaw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*. Thousand Oaks, CA: Sage.

Grissom, R.J. & Kim, J.J. (2005). *Effect Sizes for Research. A Broad Practical Approach.* Mahwah, NJ: LEA.

Hess, M. R. & Kromrey, J. D. (2004) Robust Confidence Intervals for Effect Sizes: A Comparative Study of Cohen's d and Cliff's Delta Under Non-normality and Heterogeneous Variances. Paper presented at the annual meeting of the American Educational Research Association, San Diego, April 12 – 16, 2004

Hunter, J.E. & Schmidt, F.L. (2004). *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings. Second Edition.* Thousand Oaks, CA: Sage.

Krueger, J. (2001). Null Hypothesis Significance Testing. On the Survival of a Flawed Method. *American Psychologist, 56,* 1, 16-26.

Ledesma, R., Macbeth, G. & Cortada de Kohan, N. (2008). Tamaño del Efecto: Revisión Teórica y Aplicaciones con el Sistema Estadístico ViSta. *Revista Latinoamericana de Psicología, 40,* 3, 425-439.

McGraw, K. & Wong, S. (1992). A Common Language Effect Size Statistic. *Psychological Bulletin*, 111, 361-365.

Molina-Ibañez, J.G., Ledesma, R., Valero-Mora, P. & Young, F.W. (2005). A Video Tour through ViSta 6.4, a Visual Statistical System based on Lisp-Stat. *Journal of Statistical Software, 13,* 8, 1-10.

Moore, D.S. & McCabe, G.P. (1993). *Introduction to the Practice of Statistics. Second Edition.* New York: W.H. Freeman & Company.

Randolph, J. & Edmondson, R.S. (2005). Using the Binomial Effect Size Display (BESD) to Present the Magnitude of Effect Sizes to the Evaluation Audience. *Practical Assessment Research & Evaluation*, *10*, 14. Available online: http://pareonline.net/getvn.asp?v=10&n=14

Sun, S. (2008) A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology. Published master's thesis, University of Cincinnati. Available on-line: http://www.ohiolink.edu/etd/

Thalheimer, W., & Cook, S. (2002, August). *How to Calculate Effect Sizes From Published Research Articles: A Simplified Methodology.* Available online: http://work-learning.com/effect_sizes.htm.

Thompson, B. (1994). The Concept of Statistical Significance Testing. *Practical Assessment, Research & Evaluation*, *4*, 5. Available online: http://PAREonline.net/getvn.asp?v=4&n=5 .

Thompson, B. (1998). Statistical Significance and Effect Size Reporting: Portrait of a Possible Future. *Research in the Schools, 5,* 2, 33-38.

Tierney, L. (1990). *Lisp-Stat An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. NY: John Wiley & Sons.

Valera-Espín, A. & Sánchez-Meca, J. (1997) Pruebas de Significación y Magnitud del Efecto: Reflexiones y Propuestas. *Anales de Psicología, 13*, 85-90.

Wilson Van Voorhis, C.R. & Morgan, B.L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology*, *3*, 2, 43-50.

Young, F. W., Valero-Mora, P. M. & Friendly, M. (2006). *Visual Statistic: Seeing Data With Dynamic Interactive Graphics*. Hoboken, NJ: John Wiley & Sons.

Young, F.W. (1996). *ViSta: The Visual Statistics System*. UNC L.L. Thurstone Psychometric Laboratory, Research Memorandum 94-1.

Manuscript received October 2nd, 2008
Manuscript accepted March 11th, 2009.

Apendix: Installing *ViSta* and *ES-calc*

Follow the steps below to install the program:

Step 1. Download and install ViSta. As mentioned earlier, *ES-calc* functions when integrated into *ViSta*, and so for it to work, *ViSta* must first be downloaded and installed. The latest version of *ViSta* is available at the following URL: http://www.uv.es/visualstats/Book/. From this website, one may download the program's complete code as a compressed folder. Simply decompress the folder and then run the application file ViSta.exe to open the program. The ReadMe First.txt file provides a brief description of how to install *ViSta*.

Step 2. Download ES-calc (*ES-calc.lsp*). Download the program file *ES-calc* (*ES-calc.lsp)*, available at the following URL: http://www.mdp.edu.ar/psicologia/vista/.

Step 3. Load ES-calc into ViSta. Lastly, the user should open *ViSta* and execute the command "File/Load Lisp" from the main menu (see Figure 1) to load the file *ES-calc.lsp*. This will install *ES-calc* as a *ViSta* main menu option and add the ES option to the univariate analysis command.