ORIGINAL

THE IMPORTANCE OF EFFECT SIZE: A STATISTICAL EXAMPLE USING PHYSICAL CONDITION MEASUREMENTS

IMPORTANCIA DEL TAMAÑO DEL EFECTO. UNA EJEMPLIFICACIÓN ESTADÍSTICA CON MEDIDAS DE CONDICIÓN FÍSICA

Tejero-González, C.M.[1]; Castro-Morera, M.[2] & Balsalobre-Fernández, C.[3]

1 Lecturer. Departament Educación Física, Deporte y Motricidad Humana, Universidad Autónoma de Madrid (Spain). Personal Web: www.uam.es/carlos.tejero; E-mail: carlos.tejero@uam.es
2 Lecturer. Department Métodos de Investigación y Diagnóstico en Educación, Universidad Complutense de Madrid (Spain). E-mail: maria.castro@edu.ucm.es
3 Research Student, Department of Educación Física, Deporte y Motricidad Humana, Universidad Autónoma de Madrid (Spain). E-mail: carlos.balsalobre@uam.es

ABSTRACT

It is common in the field of physical activity and sports science, as well as in other scientific disciplines, to use designs that require groups to be compared in order to determine the existence of statistically significant differences. However, information regarding the magnitude of any differences found is not always provided. This work highlights the importance of combining statistical significance with values that provide information regarding the effect size. With this in mind, and in order to provide a more didactic discussion, herein we compare the physical abilities of adolescent boys and girls and estimate the *delta parameter* statistically using the *corrected Hedges' g parameter*.

KEYWORDS: effect size, Hedges' g, delta parameter, statistical significance, α and β, group comparison, physical condition.

RESUMEN

Tanto en las ciencias de la actividad física y del deporte como en otras áreas de conocimiento científico es habitual investigar con diseños que requieren comparar grupos, concluyendo sobre la existencia o no de diferencias estadísticamente significativas. Sin embargo, no siempre se informa sobre la magnitud de las diferencias encontradas. Este trabajo subraya la importancia de acompañar la significación estadística con valores que informen sobre el tamaño del efecto. Con este propósito y en aras de una argumentación didáctica, se contrasta la capacidad física de chicos y chicas adolescentes, ejemplificando estadísticamente la estimación del *parámetro delta* con la *g de Hedges ajustada.*

PALABRAS CLAVE: tamaño del efecto, *g* de Hedges, parámetro delta, significación estadística, α y β, comparación grupos, condición física.

INTRODUCTION

It is common in the field of physical activity and sports science, as well as in other scientific disciplines, to come across reports that compare the results for various groups in order to determine the existence of statistically significant differences between them as regards a specific characteristic or variable. Thus, statistically significant differences indicate the probability (*p*) that the results observed for the response or dependent variable occur as a result of the action or influence of the independent variable rather than by chance.

In this sense, statistical significance is the likelihood that the difference between the two groups may be due to a sampling accident. In other words, it is a measure of the probability that the difference observed is the same size as that which would have been obtained by chance, even when there is no difference between the two groups. However, there are some well-known problems with

the use of significance tests as the *p* value is the result of two factors, namely the size of the difference and the size of the sample. Thus, significant results can be found when the differences between the groups are large but the sample is small, and vice versa

As a result, researchers must take two possible errors, known as *type I* and *type II* errors, into account when either designing a study or interpreting its results. A type I error, also known as a *false positive*, occurs when a null hypothesis that is actually true is rejected, in other words when the researcher infers that there is a statistically significant difference when there actually isn't one. Similarly, a type II error, also known as a *false negative,* occurs when a null hypothesis is accepted when it is actually false, in other words the researcher infers the absence of a statistically significant difference when one actually exists. Two indicators, namely α (*alpha*) and β (*beta*), are used to control type I and type II errors, respectively.

Perhaps the best known of these is the *alpha level*, or *statistical significance*, which indicates the degree of type I risk assumed by the researcher. The scientific community has established two standards for alpha: α=0.05, to perform estimations with a maximum type I error margin of 5% (95% confidence level), and α=0.01, for an error margin of 1% (99% confidence level).

However, the *beta level*, which indicates the risk assumed by the researcher that a type II error, or false negative, may occur, is also relevant. As for alpha, the scientific community has also established two standards for beta: β=0.10, when a maximum type II error margin of 10% needs to be ensured (contrast confidence or *power* of 90%), and β=0.20, to ensure a margin of error of 20% (contrast confidence or *power* of 80%).

One of the possible ways to optimise confidence levels α and β is to increase the sample size (Cañadas, Borges, Sánchez and San Luis, 2000). However, neither of these indicators provides information regarding the magnitude or importance of any differences found. As such, the researcher must use another type of indicator known as the *effect size* or *magnitude of the difference* (Fan, 2001; Frías, Pascual and García, 2000; Monterde, Pascual and Frías, 2000; Thompson, 2006; Thomas and Nelson, 2007; Valera and Sánchez, 1997).

Thus, the effect size indicates the calculated efficacy between the various levels of the independent variable, thereby complementing the information provided by the null hypothesis occurrence probability, as it provides information regarding the magnitude of any differences found as well as confirming their existence.

The work of Dowson (2000), who studied the effect of the *study moment* variable at two levels (morning and afternoon) on the *learning* variable is good example of use of the effect magnitude. Thus, this study was performed with an incidental sample of 38 individuals, who were randomly distributed into morning and afternoon groups and presented with the same learning stimulus. Their

understanding of the text was then assessed on the basis of the number of correct answers provided (maximum of 20). The mean scores were 15.2 for the morning group and 17.9 for the afternoon group. These findings raise two substantive questions, namely are the differences between the two groups sufficiently large and can it be concluded that more is learned in the afternoon than in the morning. As well as the statistical significance, one means of solving this type of question is to use the effect size. Thus, if there is no overlap between the distributions of the two groups, the difference would be important, whereas if the overlap were larger, the difference between the groups would be less important.

In order to represent this reasoning graphically, Figure 1 shows two situations in which the importance of the difference varies with the overlap of the distributions, with the differences on the left being large and significant but those on the right being less relevant.
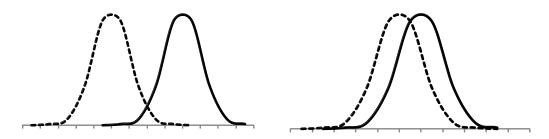

FIGURE 1: Illustration of the effect magnitude: distribution overlaps.

Various procedures can be used to estimate the effect size, including, but not limited to, the coefficient of determination, eta squared, omega squared, Phi, etc. (Rosnow, Rosenthal and Rubin, 2000; Sink and Stroh, 2006; Trusty, Thompson and Petrocelli, 2004; Vacha and Thompson, 2004). However, in order to demonstrate exactly what the effect size is, and to help the reader understand its utility and relevance, herein we will concentrate on the *standardised mean difference* or *delta parameter* (hereafter δ), which is obtained using the *corrected Hedges g* parameter (hereafter g_{corr}), following the recommendations of Ledesma, Macbeth and Cortada de Kohan (2008), rather than reviewing all the methods available for estimating effect size. This decision was taken on the basis that three factors favour the use of g_{corr}: (1) it provides and accurate and unbiased estimate, (2) it is simple to calculate, and (3) interpretation of its results is easy.

In order to obtain g_{corr}, g must first be calculated and then corrected. The g parameter is obtained from:

$$g = \frac{\left| \overline{X}_1 - \overline{X}_2 \right|}{\sqrt{\dfrac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}}}$$

where $\overline{X}_1$ is the arithmetic mean of group 1, $\overline{X}_2$ the arithmetic mean of group 2, n_1 the sample size for group 1, S_1^2 the variance in the scores for group 1, n_2 the sample size for group 2 and S_2^2 the variance in the scores for group 2.

g is then corrected as follows:

$$g_{ajust} = g \left[1 - \frac{3}{4gl - 1} \right]$$

where:

$$gl = n_1 + n_2 - 2$$

In short, g_{corr} estimates the difference between the means for the two groups and standardises it by dividing by the unified standard deviation for them, thus meaning that this procedure provides a standard parameter (z score) which is finally corrected to eliminate the sample size bias. Thus, this parameter expresses a standard value that is of great utility as it allows the percentage of cases in which one group is below the average of the other group to be inferred from the normal curve table. However, it should be noted that the normality and homoscedasticity assumptions must be fulfilled, especially for small sample sizes (less than 30 observations per group, for example; Pardo and San Martín, 2004).

In light of the above, this paper is intended to highlight the importance of accompanying the statistical significance with values that indicate the effect or magnitude of the differences. To this end, we have undertaken a group-comparison study specifically designed for this purpose that has not been previously published elsewhere.

METHOD

Participants

The sample consisted of 271 participants, 52% of whom were males (group 1; n_1= 142) aged between 12 and 18 years (M=14.44, SD=1.52), with the remaining 48% being females (group 2; n_2= 129) in the same age range (M=14.46; SD=1.52). Participants were selected by non-random incidental sampling for ease of access. Participation was voluntary, approved and unpaid.

Design, objective and variables

A retrospective *ex post facto* design. The objective of the study was to determine whether adolescent males and females differed in terms of physical condition by measuring three dependent variables, namely strength, speed and flexibility.

Hypothesis

First. Men are stronger than women.

1st-H$_0$: $\overline{X}$ strength men $\leq$ $\overline{X}$ strength women
1st-H$_1$: $\overline{X}$ strength men $>$ $\overline{X}$ strength women

Second. Men are faster than women.

2nd-H$_0$: $\overline{X}$ speed men $\leq$ $\overline{X}$ speed women
2nd-H$_1$: $\overline{X}$ speed men $>$ $\overline{X}$ speed women

Third. Men are less flexible than women.

3rd-H$_0$: $\overline{X}$ flexibility men $\geq$ $\overline{X}$ flexibility women
3rd-H$_1$: $\overline{X}$ flexibility men $<$ $\overline{X}$ flexibility women

Procedure

Data were collected by samplers with a degree in physical activity and sports science at a secondary school in the Autonomous Community of Madrid (Spain) during several physical education classes. The School Council, which consists of representatives elected by students, parents/guardians and teachers, authorised the performance of this study.

The *strength* variable was assessed by means of a *medicine ball throw* test (Legido, Segovia and Ballesteros, 1995), using a 2 kg ball and defining the strength in terms of distance thrown (in metres). To measure the *speed* variable, the participants were asked to run a distance of 50 m (Rosandich, 1999), with the time in seconds used to express the speed. Likewise, the *flexibility* variable was assessed by bending the upper body forwards from a sitting position (*Sit and Reach* test; Eurofit, 1993), using a flexibility bench equipped with a scale graduated in centimetres, with 0 being located at the soles of the feet.

Once all data had been collected and analysed, all participants received an individual report regarding their physical condition.

Data analysis

Data were analysed using inferential statistics for comparing groups, with initial confidence levels α and β of 0.05 and 0.10, respectively. All analyses were performed using the program *IBM SPSS Statistics 18*.

RESULTS

Normality and homoscedasticity assumptions

All variables must fulfil the normality assumption when comparing groups using Student's t-test and estimating the effect size, especially when comparing small groups, which is not the case here.

The results of the Kolmogorov-Smirnov test with no Lilliefors significance correction showed that the normality assumption for the strength, speed and flexibility distributions in women ($p=0.20$, $p=0.93$ and $p=0.40$, respectively), and those for the strength and flexibility in men ($p=0.74$ and $p=0.38$, respectively), should not be rejected. However, the normality hypothesis for the speed variable in men ($p=0.03$) had to be rejected. In light of the size of the group affected, and as the data for the other variables were compatible with the normal distribution hypothesis, this did not affect the aims of this study.

As far as the *homoscedasticity* or equality of variances assumption was concerned, the *Levene test* confirmed that men and women had similar dispersities for the variables speed ($F=0.199$; $p=0.65$) and flexibility ($F=0.060$; $p=0.80$), whereas this was not the case for strength ($F=53.4$; $p^< 0.001$). However, this finding does not affect our estimation of the effect size as it only occurs for one of the three variables and both groups are large and unbalanced (142 males and 129 females).

Means comparison

The *Student t-test for two independent samples* performed to compare the dependent variables for the two groups gave alpha values that allowed the three null hypotheses proposed to be rejected. In other words, it can be inferred, with a confidence level of 99%, that it is false that men are equally or less strong than women ($t=11.87$; $gl=213.13$; $p^< 0.001$), that men are equally as fast, or slower, than women ($t=-10.57$; $gl=269$; $p^< 0.001$), or that men or equally or more flexible than women ($t=-5.61$; $gl=269$; $p^< 0.001$).

This empirical evidence allows us to provisionally conclude that men are stronger and faster than women, whereas women are more flexible than men (Table 1).

TABLE 1. t-Test for comparing independent means

	Gender	M	SD	p
Strength: metres a 2 kg ball is thrown	Men	7.33	1.85	˂ 0.001
	Women	5.23	0.94	
Speed: seconds in 50 m	Men	8.29	1.06	˂ 0.001
	Women	9.60	0.96	
Flexibility: centimetres bending upper body when seated	Men	1.85	7.8	˂ 0.001
	Women	7.26	7.9	

Effect size

Estimation of the effect size using g_{corr} by solving the equation described in the introduction, gives the following *delta* magnitudes: $\delta_{strength} = 1.39$, $\delta_{speed} = 1.27$ and $\delta_{flexibility} = 0.68$.

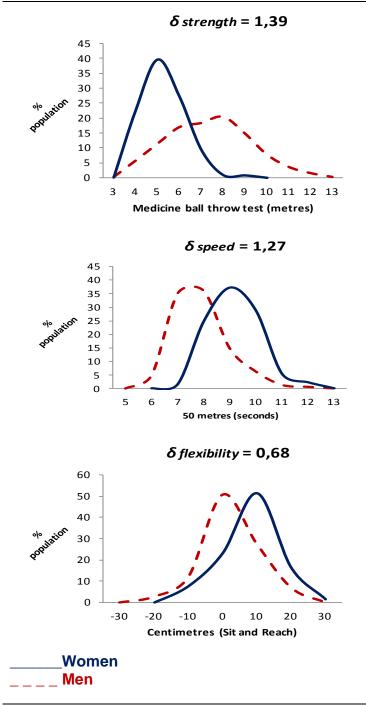These statistics correspond to the graphs contained in Figure 2.

FIGURE 2: Effect sizes and graphs

A subsequent analysis of the cumulative probability of each difference found using a *standardised normal distribution table* (see, for example, Vincent, 2005) led to three new results: 91% of women have an equal or lower strength than the average for men (z=1.39), 89% of women have an equal or lower speed than the average for men (z=1.27), and 75% of men have an equal or lower flexibility than the average for women (z=0.68; Table 2).

TABLE 2. Standardised normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
...	...	...	...	...	...	...	...	...	...	...
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

DISCUSSION AND CONCLUSIONS

This work, which is based on two statistically comparable groups in terms of size and age distribution, has shown that men and women have different physical abilities (strength, speed and flexibility) during adolescence. To reach this conclusion, the two groups have been compared using *Student's t-test for independent samples* for the three dependent variables, with the same statistical significance, namely p<0.01, being obtained in each case. In other words, the probability that a *false positive* has occurred is less than 1% for all three variables.

Up to this point, however, it can only be concluded that men and women differ in terms of these variables and that the confidence level in this statement is greater than 99%. It is also known that men are significantly stronger and faster than women and that women are statistically more flexible than men, but we don't know the size of these differences. Thus, we do not know the importance or magnitude of the differences found. To resolve this question, the

standardised mean difference or *delta parameter* has been estimated using g_{corr} and the following values found: $\delta_{strength}$=1.39, δ_{speed}=1.27, $\delta_{flexibility}$=0.68. These values allow us to infer, using the normal distribution table, that only 9% of women are stronger than the average for men, only 11% of women are faster than the average for men and only 25% of men are more flexible than the average for women.

As a result, the δ values obtained, irrespective of the scales used to measure them (the variables analysed were measured in metres, seconds and centimetres) indicate that the largest difference between groups occurred for the strength variable, followed by speed and, finally, flexibility. This information may be relevant for both theoretical and practical purposes and is not provided by the statistical significance.

Therefore, as can be seen from the results described herein, and in agreement with other authors (Fernández-Cano and Fernández-Guerrero, 2009; Lustig and Trauser, 2004; Rhea, 2004; Smith and Honoré, 2008; Thompson, 1999), the effect size has many advantages. The most important of these are listed below:

1. it is relatively easy to calculate;

2. as an expression of standard deviation or explained variance, effect size provides an intuitive interpretation of the results and the magnitude of the differences, and as such can be considered to be indicative of practical significance whilst not being incompatible with statistical significance;

3. as a result of its dimensionless nature, it allows differences between variables with different units to be compared;

4. it is a statistic that allows more knowledge to be accumulated by allowing the meta-analytical comparison of results from different studies with the same objective; and

5. it maintains a greater independence as regards the effect of sample size than statistical significance.

In short, the aim of this work was to highlight, using statistical arguments and examples, the advisability of accompanying statistical significance probabilities with other values that provide information regarding the effect size or magnitude of the differences. If this is not done, we strongly believe that the resulting report will face limitations as regards the presentation of its conclusions.

LITERATURE REFERENCES

Cañadas, I., Borges, A., Sánchez, A. y San Luis, C. (2000). Estudio de la potencia de los contrastes de medias con dos y tres grupos con tamaño del efecto pequeño y en condiciones de normalidad y homo-heterocedasticidad, *Psicothema, 12*(2), 114-116.

Dowson V. (2000). Time of day effects in school-children's immediate and delayed recall of meaningful material. *TERSE Report* (en http://www.cem.dur.ac.uk/ebeuk/research/terse/library.htm).

Eurofit (1993). *Eurofit Tests of Physical Fitness*, 2nd Edition. Strasbourg.

Fan, X. (2001). Statistical significance and effect size in education research: two sides of a coin. *The Journal of Educational Research, 94*(5), 275-282.

Fernández-Cano, A. y Fernández-Guerrero, I. (2009). *Crítica y alternativas a la significación estadística en el contraste de hipótesis*, Madrid: La Muralla.

Frías, M. D., Pascual, J. y García, J. F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema, 12*(2), 236-240.

Ledesma, R., Macbeth, G. y 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.

Legido, J. C., Segovia, J. C. y Ballesteros, J. M. (1995). *Valoración de la condición física por medio de test*. Madrid: Ediciones pedagógicas.

Lustig, D. y Trauser, D. (2004). Effect size and rehabilitation research. *Journal of Rehabilitation, 70*(4), 3.

Monterde, H., Pascual, J. y Frías, M. D. (2000). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones, *Psicothema, 18*(4), 848-856.

Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II*, Madrid: Pirámide.

Rhea, M. (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *Journal of Strength & Conditioning Research, 18*(4), 918-920.

Rosandich, T. P. (1999). International Physical Fitness Test. *The Sport Journal*, *2*(1).

Rosnow, R. L., Rosenthal, R. y Rubin, D. B. (2000). Contrasts and correlation in effect-size estimation. Psychology Science, 11(6), 446-453.

Sink, A. y Stroh, H. R. (2006). Practical significance: the use of effect sizes in school counseling research. *Professional School Counseling Journal, 8*(1), 115-120.

Smith, M. L. y Honoré, H. H. (2008). Effect size reporting in current health education literature. *American Journal of Health Studies, 23*(3), 130-135.

Thomas, J. R. y Nelson, J. K. (2007). *Métodos de investigación en actividad física*. Badalona: Editorial Paidotribo.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. *The Journal of Psychology, 133*(2), 133-140.

Thompson, B. (2006). Research synthesis: effect sizes. En J. Green, G. Gamilli y P. B. Elmore (Eds.). Handbook of complementary methods in education

research (pp. 583-603). Washington, DC: American Educational Research Association.

Trusty, J., Thompson, B. y Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the Journal of Counselling and Development. *Journal of Counselling & Development,* 82, 107-110.

Vacha, T. y Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*(4), 473-481.

Valera, A. y Sánchez, J. (1997). Pruebas de significación y magnitud del efecto: Reflexiones y propuestas, *Anales de Psicología, 13*(1), 85-90.

Vincet, W. J. (2005). Statistics in Kinesilogy (3rd ed.). Human Kinetics.

Número de citas totales / Total references: 23 (100%)

Número de citas propias de la revista / Journal's own references: 0 (0 %)