Calcular e apresentar tamanhos do efeito em trabalhos científicos (1): As limitações do $p < 0,05$ na análise de diferenças de médias de dois grupos

Calculating and reporting effect sizes on scientific papers (1): $p < 0.05$ limitations in the analysis of mean differences of two groups

Artigo de Revisão | Revision Article

Helena Espírito-Santo, PhD (1a), Fernanda Daniel, PhD (1a)

(1) Instituto Superior Miguel Torga, Coimbra, Portugal.
(2) Instituto Superior Miguel Torga, Coimbra, Portugal.
(a) Elaboração do trabalho.

Autor para correspondência | Corresponding author: Helena Espírito-Santo; Largo de Celas, 1, 3000-132 Coimbra, Portugal; +351 910637946; helenum@gmail.com.

Palavras-Chave
Tamanho do efeito
Significância estatística
Valor *p*
d de Cohen
g de Hedges
Delta de Glass

RESUMO

A *Revista Portuguesa de Investigação Comportamental e Social* exige que os autores sigam as recomendações do *Publication Manual of the American Psychological Association* (APA, 2010) na apresentação da informação estatística. Uma das recomendações da APA é de que os tamanhos do efeito sejam apresentados associados aos níveis de significância estatística. Uma vez que os valores de *p* decorrentes dos resultados dos testes estatísticos não informam sobre a magnitude ou importância de uma diferença, devem então reportar-se os tamanhos do efeito (TDE). De facto, os TDE dão significado aos testes estatísticos, enfatizam o poder dos testes estatísticos, reduzem o risco de a mera variação amostral ser interpretada como relação real, podem aumentar o relato de resultados "não-significativos" e permitem acumular conhecimento de vários estudos usando a meta-análise.

Assim, os objetivos deste artigo são os de apresentar os limites do nível de significância; descrever os fundamentos da apresentação dos TDE dos testes estatísticos para análise de diferenças entre dois grupos; apresentar as fórmulas para calcular os TDE, fornecendo exemplos de estudos nossos; apresentar procedimentos de cálculo dos intervalos de confiança; fornecer as fórmulas de conversão para revisão da literatura; indicar como interpretar os TDE; e ainda mostrar que, apesar de frequentemente ser interpretável, o significado (efeito pequeno, médio ou grande para uma métrica arbitrária) pode ser impreciso, havendo necessidade de ser interpretado no contexto da área de investigação e de variáveis do mundo real.

Keywords
Effect size
Statistical significance
p-value
Cohen's *d*
Hedges' *g*
Glass's Delta

ABSTRACT

The Portuguese Journal of Behavioral and Social Research requires authors to follow the recommendations of the Publication Manual of the American Psychological Association (APA, 2010) in the presentation of statistical information. One of the APA recommendations is that effect sizes should be presented along with levels of statistical significance. Since p-values from the results of the statistical tests do not indicate the magnitude or importance of a difference, then effect sizes (ES) should reported. In fact, ES give meaning to statistical tests; emphasize the power of statistical tests; reduce the risk of interpret mere sampling variation as real relationship; can increase the reporting of "non-significant"results, and allow the accumulation of knowledge from several studies using meta-analysis.

Thus, the objectives of this paper are to present the limits of the significance level; describe the foundations of presentation of ES of statistical tests to analyze differences between two groups; present the formulas to calculate directly ES, providing examples of our own previous studies; show how to calculate confidence intervals; provide the conversion formulas for the review of the literature; indicate how to interpret the ES; and show that, although interpretable, the meaning (small, medium or large effect for an arbitrary metric) could be inaccurate, requiring that interpretation should be made in the context of the research area and in the context of real world variables.

Recebido | Received: 04/01/2015
Revisto | Reviewed: 01/02/2015
Aceite | Accepted: 10/02/2015

VISÃO GERAL

O poder de um teste estatístico corresponde à probabilidade de rejeitar corretamente a hipótese nula (Cohen, 1992a, p. 98) e depende de três aspetos, incluindo o tamanho do efeito/TDE, o nível de significância e o tamanho da amostra (Cohen, 1988, p.4; Cohen, 1992a, p. 98; Cohen, 1992b, pp. 99-100). No presente artigo abordaremos somente o nível de significância e o tamanho do efeito.

Nível de Significância

O nível de significância corresponde à evidência de que o fenómeno existe ou ao risco de rejeitar erradamente a hipótese nula (Cohen,1988; 1992b). Apesar de dominar a literatura científica, este parâmetro não autoriza a fazer qualquer afirmação sobre a probabilidade matemática de determinada hipótese (Fisher, 1959).

História e definições. Quando, em 1925, Fisher defendeu o cálculo do valor da probabilidade (*p*) da estatística e Neyman e Pearson em 1933 apresentaram o conceito de nível de significância, ou alfa, todos estariam longe de pensar que os seus conceitos iriam dominar a literatura científica.

Para Fisher (1925; 1959), o valor de *p* de um teste estatístico mede a força da Evidência contra a Hipótese nula ($p(E|H)$). O autor defendeu um valor de *p* inferior a 0,05 como um "valor suficientemente pequeno", mas não como uma regra de ouro.

No método de Neyman e Pearson, o valor de alfa associa-se ao nível de confiança com que o teste estatístico permite rejeitar a hipótese nula (H_0 = não há um determinado resultado), tendo sido acordados os valor fixos de 0,05 (i. e., os resultados são "significativos", havendo 5% de probabilidade de obter os dados se a H_0 fosse verdadeira)[i], 0,01 (i. e., os resultados são "altamente significativos") ou 0,001 (i. e., os resultados são "muito altamente significativos"). Neste método, há que distinguir o erro do Tipo I (rejeição falsa de H_0) e o erro do Tipo II (manutenção falsa de H_0). A probabilidade do erro do Tipo I é alfa, e a probabilidade do erro do Tipo II é beta. O poder é o complemento de beta (1 - β), definindo-se como a probabilidade de rejeitar corretamente a H_0 quando a hipótese alternativa é verdadeira (H_1 = há um determinado resultado) (Cohen, 1992a; Kline, 2004).

Mais tarde, Cohen (1992b, p. 99) sugeriu que um *p* inferior a 0,05 devia ser usado para estudos exploratórios e que valores de alfa mais pequenos deviam ser usados quando se testam várias hipóteses nulas.

Desde então, a maioria dos investigadores normalmente espera obter um valor de *p* suficientemente pequeno (i. e., $p < \alpha$), para que possa rejeitar a hipótese nula, aceitando a hipótese alternativa (Huberty, 1993). Estima-se que cerca de 90% dos artigos nas ciências sociais empreguem este procedimento (Loftus, 1991), independentemente se são testes *t*, análises de variância ou estatísticas não-paramétricas, como os teste *U* de Mann-Whitney ou o teste de Kruskal-Wallis para análise de variância.

Críticas. O debate sobre a importância do nível de significância não é novo (Cumming, 2012; Kirk, 1996; Salsburg, 2002). Em 1900, Karl Pearson já afirmava que o *nível de significância* não é suficiente por si mesmo e desde a década de 60 do século XX que se registam revisões sobre a questão fundadora do valor do *p* (p. e., Morrison e Henkel 1970; Nickerson, 2000), incluindo tentativas para corrigir a situação (p. e., Giere, 1972; Trafimow, 2003), propostas para estatísticas ou métodos alternativos (p. e., Carver, 1978; Loftus, 1996)[ii] ou mesmo ataque aos níveis de significância (p. e., Cohen, 1994; Killeen, 2005).

De facto, as críticas aos limites dos níveis de significância estão a generalizar-se em várias áreas de investigação, incluindo, por exemplo, a psicologia (Loftus, 1996; Wagenmakers, 2007), neurociências (Hentschke e Stüttgen, 2011), biologia (Nakagawa e Cuthill, 2007), medicina (Aickin, 2004; Snyder e Lawson, 1993), educação (Kirk, 1996; Olejnik e Algina, 2000), desporto (Thomas, Salazar e Landers, 1991) ou o marketing (Fern e Monroe, 1996).

Importa, então, referir algumas das limitações do nível de significância. Note-se, primeiramente, que o teste de significância da hipótese nula é assimétrico, pelo que só se podem reunir dados *contra* a hipótese nula, mas não em seu suporte. Assim, os valores de *p* acima de 0,05 não indicam a ausência de um efeito (revisão de Ferguson, 2009; Hentschke e Stüttgen, 2011; Loftus, 1996).

Assinale-se também que, apesar de o valor do *p* fornecer a probabilidade de se obter uma estatística significativa, ele não informa sobre a importância clínica ou prática dos resultados (Berben, Sereika e Engberg, 2012; Durlak, 2009; Jacobson e Truax, 1991; Kirk, 1996; Snyder e Lawson, 1993), sendo somente adequado se se explicitar o seu papel no suporte a uma teoria (Chow, 1988). Na realidade, só se devia dizer que um *p* é menor do que alfa quando a hipótese fosse colocada *antes* dos dados serem recolhidos (Kline, 2004).

O nível de alfa é também criticado pela sua arbitrariedade (p. e., Kline, 2004), com Rosnow e Rosenthal (1989, p. 1277) a escreverem ironicamente "...

God loves the 0.06 nearly as much as the 0.05"[iii]; com Cochran, Moses e Mosteller (1983, p. 19) a afirmarem "no strong logical reason lies behind this choice"[iv], e ainda com Kirk (1996, p. 748) a criticar que o investigador passa de um *continuum* de incerteza para uma decisão dicotómica de rejeitar-não-rejeitar (respetivamente: p = 0,06 ou p < 0,05)[v].

Por estas razões, há autores que aconselham a especificar um nível de alfa que considere a *importância relativa* do erro do Tipo I *versus* o erro do Tipo II, decidindo refletidamente qual a *probabilidade aceitável* de encontrar um *falso* efeito (erro do Tipo I) e a probabilidade aceitável de não detetar um efeito (erro do Tipo II; Aguinis et al., 2010, pp. 520-521). Depois, ao escolher um alfa no nível de 0,05 ou menor, há que ter em conta que a *importância relativa* do erro do Tipo I *versus* o erro do Tipo II varia consoante as áreas de investigação, os resultados e os contextos, podendo ser menos arriscado encontrar um falso efeito do que perder um efeito que existe e que faz a diferença para a sociedade (Aguinis et al., 2010).

Erros comuns. Acauteladas as limitações do valor *p*, depois, na apresentação do nível de significância há que ter em atenção em não seguir quatro erros comuns. Quando se verifica, por exemplo, que a diferença entre as médias dos dois grupos não é estatisticamente significativa, isto não quer dizer que "não há diferença", mas sim que não há evidência para *rejeitar a hipótese nula de que não há diferenças* (Cohen, 1988). Na verdade, pode dizer-se que "não há diferença" quando o *poder* do teste é alto[vi]. Nas palavras de Ellis (2010), o mais correto é dizer que o resultado é "inconclusivo" (p. 53).

O segundo erro é acreditar que um valor de *p* igual a, por exemplo, 0,01, significa que se a experiência fosse replicada, obter-se-ia um resultado significativo em 99% das repetições (Carver, 1978, p. 392).

Quanto ao terceiro erro, há que não confundir a significância estatística com a significância clínica ou científica (Berben et al., 2012; Durlak, 2009; Jacobson e Truax, 1991). Ao ler-se num artigo que "houve uma melhoria significativa no funcionamento cognitivo", o que quer dizer a palavra "significativa"? Quer dizer que a melhoria foi grande, importante ou que o *p* foi inferior a 0,05? Ora, um pequeno *p* não garante que a melhoria seja grande ou importante (ver tópico seguinte). Por isso, Kline (2004) recomenda que se deixe cair a palavra "significante" e se substitua "por estatística", devendo então a frase em cima reescrever-se como: "houve uma melhoria estatística no funcionamento cognitivo".

Finalmente, em relação ao quarto erro, como se terá reparado acima, o valor de *p* é $p(E|H)$, no entanto, a maior parte dos investigadores lê o 0,05 como $p(H|E)$, isto é, como a probabilidade de a Hipótese ser verdadeira, dada a Evidência (Carver, 1978; Kirk, 1996). Ora, e dito de outro modo, a probabilidade de que é um falante português ao estar a ler este artigo (próxima de 1) é diferente da probabilidade de ler este artigo por ser um falante de português (adaptado de Cumming, 2012). O modo irónico como Carver (1978) o expressa é mais ilustrativo:

> What is the probability of obtaining a dead person (label this part D) given that the person was hanged (label this part H); this is, in symbol form, what is $p(D|H)$? Obviously, it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has hanged (H), given that the person is dead (D); that is, what is $p(H|D)$? This time the probability will undoubtedly be very low, perhaps 0.01 or lower.[vii] (Carver, 1978, p. 393).

Valor de *p* vs. tamanho do efeito. O nível de significância é afetado por, pelo menos, sete características dos estudos (Schneider e Darcy, 1984), sendo o tamanho da amostra a mais determinante (Snyder e Lawson, 1993). Assim, é mais provável obter um valor de *p* significativo com tamanhos grandes das amostras e, inversamente, em amostras pequenas, o valor de *p* pode não ser significativo, ainda que o TDE possa ser grande (Berben et al., 2012; Durlak, 2009; Snyder e Lawson, 1993). Na realidade, todo e qualquer teste estatístico pode ser expresso como o produto (Rosenthal, 1994):

Teste estatístico = TDE x *função* do tamanho da amostra.

Assim, um desejável valor pequeno de *p* pode associar-se a um TDE pequeno, a um médio e também a um grande (Berben et al., 2012; Durlak, 2009; Rosnow e Rosenthal, 1989). Snyder e Lawson (1993) mostraram que o acrescento de um só sujeito, num estudo com uma pequena amostra e com um grande TDE de 0,66, o valor *p* descia abaixo de 0,05 sem que a magnitude do TDE se alterasse.

Para melhor compreensão destes aspetos, na Tabela 1, vejam-se quatro estudos hipotéticos com comparações entre as médias de dois grupos e atente-se aos valores de *p*.

Como se pode ver, a dimensão da diferença foi muito maior no estudo B do que no estudo A, apesar de ambas as diferenças serem significativas e os graus de liberdade serem os mesmos. Depois, note-se como a dimensão da diferença foi maior no estudo C do que no D, apesar de ambas as diferenças não serem significativas. Repare-se ainda como a magnitude da diferença no estudo A foi

Tabela 1
Comparação das Médias de Dois Grupos em Quatro Estudos Hipotéticos

Estudo	*gl*	Grupo alvo	Grupo de controlo	M_1 - M_2	*t*	*p*
		$M_1 \pm DP_1$	$M_2 \pm DP_2$			
A	34	6,50 ± 1,27	3,70 ± 1,77	2,80	4,07	0,001
B	34	2,00 ± 1,49	9,00 ± 1,49	7,00	10,50	< 0,001
C	34	14,78 ± 1,09	19,9 ± 8,58	5,12	1,87	0,078
D	34	6,00 ± 1,49	5,00 ± 1,49	1,00	1,50	0,151

Nota: gl = graus de liberdade; M = média, DP = desvio-padrão.

menor do que no estudo C, apesar de a diferença ser significativa no estudo A e não ser significativa no estudo C.

Em síntese, se um investigador usar somente o nível de significância, o valor da magnitude da diferença pode perder-se, sendo uma perda especialmente grave se a investigação disser respeito à avaliação da eficácia de um programa de intervenção (Rosenthal, 1983).

Necessidade e Interesse em Apresentar os Tamanhos do Efeito

Todos os testes estatísticos têm o seu próprio índice de TDE (Cohen, 1992b). Estes índices do TDE, para além de terem a vantagem de não dependerem do tamanho da amostra[viii], informam sobre o significado dos resultados e consistem numa métrica comum para comparar resultados de estudos diferentes (Berben et al., 2012; Bezeau e Graves, 2001; Cumming, 2012; Ferguson, 2009; Kline, 2004; Lipsey et al., 2012; Olejnik e Algina, 2000; Snyder e Lawson, 1993).

De facto, há mais de 20 anos que na literatura científica se chama repetidamente a atenção para a necessidade de comunicar os tamanhos de efeito (p. e., Cohen, 1990, 1994). Wilkinson e a *American Psychological Association Task Force on Statistical Inference*, em 1999, tinham já publicado diretrizes para a pesquisa publicada na psicologia declarando que se deve "always present effect sizes for primary outcomes"[ix], enfatizando que "reporting and interpreting ESs [effect sizes] in the context of previously reported effects is essential to good research"[x] (Wilkinson Task Force on Statistical Inference, 1999, p. 599). A sexta edição do *Publication Manual* da APA (Association, 2010) também destaca a importância de incluir os TDE e os intervalos de confiança quando se aplicam estatísticas inferenciais.

Apesar das recomendações, o problema encontra-se nas instruções das revistas e na publicação dos artigos. Assim, na nossa revisão contabilizámos 609 publicações nas ciências sociais e comportamentais indexadas no *Journal Citation Reports* do ISI-Thompson, Scielo e Scopus, englobando as áreas do desporto (*n* = 11), educação e psicologia educacional (*n* = 20), etologia e ecologia comportamental (*n* = 16), gerontologia (*n* = 30), neurociências (*n* = 93), neuropsicologia (*n* = 59), psicofarmacologia e comportamentos aditivos (*n* = 28), psicologia (*n* = 130, incluindo, geral, social, forense, organizacional e desenvolvimental, etc.), psiquiatria e saúde mental, familiar e sexual (*n* = 172), e em outras ciências sociais (*n* = 50, incluindo antropologia, epidemiologia, serviço social e sociologia), e verificámos que somente 58 recomendam explicitamente o cálculo da magnitude do efeito (9,5%).

Ao nível das publicações de trabalhos, já Cohen (1962), na sua revisão de publicações numa revista de psicologia, mostrou que a omissão na apresentação dos TDE podia conduzir à falha na rejeição da hipótese nula. Outras revisões têm também mostrado a mesma negligência noutras áreas de investigação (p. e., Bezeau e Graves, 2001; Caperos e Pardo, 2013; McMillan e Foley, 2011; Schatz, Jay, McComb e McLaughlin, 2005; Sedlmeier e Gigerenzer, 1989; Sun, Pan e Wang, 2010).

DEFINIÇÃO E MÉTODOS DE CÁLCULO DOS TAMANHOS DO EFEITO

As primeiras e mais citadas definições dos TDE incluem o "grau em que a hipótese nula é falsa" ou, ainda, o "grau em que um fenómeno está presente na população" (Cohen, 1988, pp. 9-10).

No entanto, na maioria das vezes, os dados relativamente à população não estão disponíveis[xi], pelo que o TDE se refere à estimativa da magnitude da relação entre variáveis, do efeito de uma variável sobre outra, ou da diferença entre duas amostras (Hedges, 1981; Rosenthal, 1994).

A este respeito, existem duas famílias principais de TDE, a família dos TDE baseados nas médias e desvios-padrão, em que o TDE é a expressão da magnitude da diferença entre grupos em relação a uma determinada característica; e a família respeitante à magnitude da relação entre duas variáveis. Neste artigo, vamos abordar a primeira família, ficando o TDE das relações para um próximo artigo.

Tamanhos do efeito baseados nas médias e desvios-padrão

O TDE pode ser não-padronizado ou padronizado. A diferença simples entre duas médias pertence à magnitude *não-padronizada*. Este tipo de magnitude mantém-se ligada à unidade de medida (p. e., número de experiências traumáticas; pontuação numa escala de solidão) e pode ser preferível se o investigador retirar um significado intuitivo da unidade de medida (Hentschke e Stüttgen, 2011). Baguley (2009), considera mesmo que a magnitude não-padronizada é mais robusta e versátil do que a padronizada.

Nos estudos com delineamento por grupos, pode ser adequada a análise da diferença *padronizada* da média, pois é independente da métrica. A magnitude padronizada é dimensionada em termos de variabilidade da amostra ou população de onde a medição foi efetuada (Baguley, 2009). Consoante se procura a diferença padronizada entre dois grupos ou entre mais do que dois grupos, assim se tem diferentes opções de se efetuar os cálculos. Neste artigo, vamos abordar somente as comparações entre dois grupos.

Comparações Entre Dois Grupos. Quando se está perante dois grupos, a análise da diferença padronizada entre médias pertence à família *d* que inclui o *d* de Cohen, o *g* de Hedges e o delta de Glass. O numerador é o mesmo nos três (a diferença entre duas médias), variando o denominador para se obter a padronização (cf. equações na Tabela 2). As diretrizes indicam que, depois de se efetuarem os cálculos, há que explicitar qual o TDE usado; apresentar os efeitos para todos os resultados, independentemente de serem ou não significativos; e especificar como os TDE foram calculados fornecendo a referência apropriada ou fornecendo a equação usada (Durlak, 2009).

d de Cohen. O *d* de Cohen é uma medida comum do TDE para testes *t* com grupos independentes, ainda que seja comum haver confusão entre o parâmetro populacional proposto por Cohen (Cohen, 1988), ou índice δ, e o índice *d* propriamente dito que é uma estimativa para amostras (Borenstein, 2009). O índice δ corresponde à padronização da diferença entre médias verdadeiras (populacionais) de dois grupos e usa-se quando os *n* dos dois grupos são semelhantes e os desvios-padrão verdadeiros (populacionais) são também similares, permitindo-se a suposição de que representam uma estimação de um desvio-padrão populacional comum (Cohen, 1988; 1992a; Ellis, 2010, p. 10; Rosenthal, 1994), tal como é assumido na maior parte das técnicas paramétricas (Borenstein, 2009; Tabela 2, Equação 2.1). O índice *d* usa-se para o mesmo tipo de situação, mas com duas amostras (Tabela 2, Equação 2.2).

O *d* de Cohen é também aplicável a testes *t* para amostras emparelhadas (p. e., pares emparelhados, medidas repetidas, pré e pós intervenção num só grupo). Para o efeito, ou usa-se na mesma a Equação 2.2 (Borenstein, 2009), ou usa-se o desvio-padrão *médio* de ambos os grupos emparelhados (Equação 2.3; Cumming, 2012; Lakens, 2013). No entanto, para as situações de comparação entre a pré e pós intervenção, de acordo com Lipsey e colaboradores (2012, p. 5), ao incluir as covariáveis da linha base, a melhor estimativa do efeito da intervenção consiste na diferença das médias ajustadas pelas covariáveis (médias estimadas marginais), não

Tabela 2

Equações e Recomendações para o Cálculo dos Tamanhos do Efeito mais Comuns em Diferenças entre Médias de Grupos Independentes

Equação	Estimativa	Magnitude da diferença	Desvio-padrão		Uso
2.1	δ de Cohen	$\frac{\mu_1 - \mu_2}{\sigma}$	σ =	Estimativa do desvio-padrão populacional	*DP* populacional conhecido
2.2	*d* de Cohen	$\frac{M_1 - M_2}{DP_{combinado}}$	$DP_{\text{combinado}}$ =	$\frac{(n_1-1)DP_1^2 + (n_2-1)DP_2^2}{n_1+n_2-2}$	Grupos independentes *DP* similares
2.3	d_m de Cohen	$\frac{M_1 - M_2}{\frac{DP_1+DP_2}{2}}$	$\frac{DP_1+DP_2}{2}$ =	Média dos desvios-padrão de ambos os grupos	Grupos emparelhados
2.4	*g* de Hedges	$\frac{M_1 - M_2}{DP_{combinado}}(1 - \frac{3}{4gl-1})$	$DP_{\text{combinado}}$ =	$\frac{(n_1-1)DP_1^2 + (n_2-1)DP_2^2}{n_1+n_2-2}$	Tamanhos dos grupos diferentes; meta-análise
2.5	Δ de Glass	$\frac{M_1 - M_2}{DP_{Controlo}}$	DP_{controlo} =	*DP* do grupo de controlo	*DP* diferentes e violação da homogeneidade de variâncias

Notas: *M* = média de cada grupo; *DP* = desvio-padrão de cada grupo; *n* = número de sujeitos; *gl* = graus de liberdade (*n* -1) e d_m = desvio-padrão médio. Os números 1 e 2 em índice referem-se ao grupo 1 (p. e., grupo experimental, grupo de tratamento ou grupo alvo) e ao grupo 2 (p. e., controlo ou de comparação). Tabela baseada nas seguintes referências: Cohen (1992); Borenstein (2009); Cummings (2012); Hedges (1981); Lakens (2013) e Rosenthal (1994).

devendo ajustar-se os desvios-padrão. Nestes casos, sugere-se a análise da covariância (ANCOVA; Tabachnick e Fidell, 2007).

Nas situações em que as amostras são pequenas[xii], como o *d* de Cohen tende a ser sobrestimado, usa-se então o *g* de Hedges como fator de correção.

g de Hedges. O *g* de Hedges é assim designado em honra de Gene V. Glass que foi um dos pioneiros da meta-análise. O *g* de Hedges é recomendado quando os grupos têm pequena dimensão e são diferentes no tamanho, pelo que os desvios-padrão de cada grupo devem ser ponderados segundo a dimensão da amostra (Ellis, 2010; Hedges, 1981; Rosnow, Rosenthal e Rubin, 2000). O fator de correção *J* é aplicado na Equação 2.3 (*J* = 1-3/(4*gl*-1); Tabela 2) e deve ser sempre inferior a 1,0, pelo que o *g* será sempre inferior ao *d.* Quando os graus de liberdade são muito pequenos (p. e., < 10), a diferença entre os dois índices é mínima (Borenstein, 2009, p. 227), pelo que não é necessário proceder à correção.

Delta* de *Glass. O delta de Glass usa-se quando os desvios-padrão dos dois grupos diferem, havendo comprometimento da homogeneidade da variância (Ellis, 2010). Este índice, usado nos estudos de intervenção (Lipsey et al., 2012), consiste no valor amostral da diferença da média padronizada através do desvio-padrão do grupo de controlo (Glass, 1976; Hedges, 1981; Rosenthal, 1994; Tabela 2, Equação 2.4), com o suposto de que o desvio-padrão do grupo de controlo não é afetado pelos efeitos do tratamento, refletindo mais de perto o desvio-padrão da população (Ellis, 2010; Rosenthal, 1994). Outra alternativa, para quando os desvios-padrão dos dois grupos diferem muito, será transformar os dados por forma a tornar os desvios-padrão mais similares (p. e., transformação logarítmica, raiz quadrada; Rosenthal, 1994, p. 235) e depois proceder ao cálculo do índice *d* ou do índice *g*.

Apresentação dos intervalos de confiança. O intervalo de confiança (IC) é o intervalo de valores onde se estima que o parâmetro populacional se situe para uma dada probabilidade. Quanto mais estreito o IC, mais precisa é a estimativa (Cumming, 2012). Os ICs em relação à média devem ser apresentados por forma a aumentar a precisão de estimativas que se baseiam em amostras (Berben et al., 2012; ver p. 1041 para a fórmula do seu cálculo) e o mesmo deve ser feito em relação aos TDE (Berben et al., 2012; Breaugh, 2003; Durlak, 2009). A equação para o IC a 95% em redor do *d* é a seguinte (Hedges e Olkin, 1985, p. 86):

IC_d 95% = d ± 1,96 x *DP* de d, em que *DP* de d = $\sqrt{\frac{N}{n_1+n_2}+\frac{d^2}{2N}}$

em que 1,96 é o valor crítico da distribuição normal padronizada em 0,05 de significância; o *DP* é o desvio-padrão; o *N* representa o número total da amostra; o *n* refere-se ao tamanho da amostra de cada grupo particular (1 = grupo alvo; 2 = grupo de comparação); e o *d* é o valor do *d* de Cohen (nos suplementos, apresentamos uma folha de cálculo para este cálculo).

Enquanto que o TDE não é afetado pelo tamanho da amostra, a precisão do seu IC 95% é afetada, sendo habitualmente maior a precisão quanto maior for a amostra. Assim, para amostras pequenas há que corrigir o enviesamento através do fator de correção seguinte (Berben et al., 2012; Cooper, Hedges e Valentine, 2009):

$$J = 1 - \frac{3}{4gl - 1}$$

em que *gl* são os graus de liberdade. Depois multiplica-se IC_d 95% por *J*.

Computação rápida. Uma das razões para o nível de significância continuar a ser reportado e o TDE não, pode estar relacionado com o facto de muitos programas de análise estatística, como o SPSS, não efetuarem esse cálculo (Hentschke e Stüttgen, 2011). Este artigo, nos seus suplementos, é acompanhado por uma folha de cálculo em Excel para computação rápida desligada da Internet. A nossa folha de cálculo inclui os cálculos para o *d*, *g* e *Delta* e ainda a equivalência em percentis. Na nossa folha de cálculo fornecemos também a computação dos ICs.

Conversões para revisão da literatura. Um dos usos da magnitude do efeito aplica-se à revisão da literatura. No entanto, como já referimos, raramente os TDE são indicados na literatura. No entanto, desde que os dados básicos sejam apresentados (p. e., *n*, média e *DP*), o autor pode calcular facilmente os TDE da literatura para comparação. Se esses valores estiverem em falta, os TDE ainda podem ser calculados com base noutras estimativas (Durlak, 2009). Para esse efeito, existem fórmulas de conversão de testes estatísticos comuns em *d* de Cohen. Para os objetivos deste artigo, interessa aqui somente apresentar a conversão do teste *t* em *d* de Cohen (Rosenthal, 1994):

$$d = \frac{t\,(n_1 + n_2)}{\sqrt{gl}\sqrt{n_1 n_2}}$$

No caso de se poder considerar o tamanho das duas amostras igual ou quando não existe informação, então a equação é simplificada para (Rosenthal, 1994):

$$d = \frac{2t}{\sqrt{gl}}$$

O *g* de Hedges, quando se assume que as duas amostras têm tamanho igual corresponde a (Rosenthal, 1994):

$$g = \frac{2t}{\sqrt{n_1 + n_2}}$$

Nestas três fórmulas, o *t* corresponde ao teste t de Student, o *n* equivale ao tamanho da amostra de cada grupo e o *gl* aos graus de liberdade.

Interpretação dos Tamanhos do Efeito. Cohen (1988), hesitantemente, classificou os TDE em *pequeno*, *moderado* e *grande* (Tabela 3), escrevendo que "there is a certain risk in inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral Science"[xiii] (p. 25). Posteriormente, Cohen (Cohen, 1992b), p. 99) explanou que um TDE *moderado* representa uma magnitude evidente a olho nu para um investigador cuidadoso, um *pequeno* é claramente menor, e um *grande* é claramente maior. Uma leitura cuidadosa dos textos de Cohen mostra que, de facto, o estatístico nunca pretendeu que a sua classificação fosse aplicada de forma rígida, mas somente fornecer "some sense of its [TDE] scale"[xiv] (Cohen, 1992b, p. 99). A folha de Excel, nos suplementos, fornece esta interpretação facilmente.

Tabela 3
Valores para Interpretação dos Tamanhos do Efeito

Insignificante	Pequeno	Médio	Grande	Muito grande
< 0,19	0,20 - 0,49	0,50 - 0,79	0,80 - 1,29	> 1,30

Notas: Estes valores foram apresentados por Cohen (1988, p. 40). Rosenthal (1996) acrescentou a classificação de "muito grande".

Críticas e sugestões. Se Cohen hesitou, Glass e seus colegas (1981, p. 104) opuseram-se veementemente à adjetivação dos efeitos, considerando que estes não têm nenhum valor inerente se forem separados do contexto e de valores comparativos. Acrescente-se que um TDE maior num estudo A do que num estudo B não significa que a teoria seja mais apoiada pelo estudo A (Chow, 1988). Para além de que estas categorias não se aplicam a todas as áreas de investigação (Olejnik e Algina, 2000), podendo uma magnitude do efeito ser relevante numa área e insignificante noutra (Lipsey et al., 2012).

As classificações, ainda assim, poderão ser usadas especialmente para resultados totalmente novos e que não podem ser comparados com outros achados na literatura (Cohen, 1988), tendo sempre em consideração que quanto maior a pontuação maior o efeito. Assim, um valor de 0,05 significa que a diferença das médias dos dois grupos corresponde a metade de um desvio-padrão e um valor de 1,00 significa que a diferença das médias dos dois grupos corresponde a um desvio-padrão (Ellis, 2010, pp. 10-11).

Tamanho do efeito em linguagem comum. Outra interpretação dos TDE, menos usada, e designada por *TDE em linguagem comum* [(TDE-LC; *common language effect size statistic*; McGraw e Wong (1992)] ou por *probabilidade de superioridade* (Grissom e Kim, 2005), é mais intuitiva, pois converte o efeito numa percentagem, podendo ser calculada a partir do *d* de Cohen (Lakens, 2013). Esta percentagem associa-se à probabilidade no extremo superior do valor de *Z* da equação seguinte:

$$Z = \frac{|M_1 + M_2|}{\sqrt{\frac{DP_1^2 + DP_2^2}{2}}}$$

Nesta equação, o *M* corresponde às médias de cada grupo e o *DP* aos desvios-padrão também de cada grupo.

O TDE-LC corresponde à probabilidade de um valor de *Z* ser maior do que o valor que corresponde a uma diferença de zero entre os grupos numa distribuição normal. Dito de uma maneira mais simples para o leigo em estatística: o TDE-LC corresponde à probabilidade de uma pessoa retirada ao acaso do grupo alvo ter uma pontuação superior a uma pessoa retirada ao acaso do grupo de comparação. Para demonstrar o valor do TDE-LC, McGraw e Wong (1992) exemplificaram com as alturas de homens e mulheres que, em média, correspondiam respetivamente a 1,77 m (*DP* = 7,11 cm) e 1,63 m (*DP* = 6,60 cm). O *d* de Cohen é igual a 2,00 e o TDE-LC é de 0,92; ou seja, a probabilidade de um homem ser mais alto do que uma mulher é de 92%, o que é mais intuitivo de que dizer que o TDE é muito grande. Dunlap (1994, p. 509), considerou este TDE-LC "an appealing index of effect size that requires no prior knowledge of statistics to understand"[xv]. Na Tabela 4, apresentamos a conversão dos TDE em probabilidades de superioridade (na folha Excel, nos suplementos, fornecemos também uma computação rápida).

O TDE-LC é também facilmente calculado através do SPSS. Primeiro computa-se o teste *U* de Mann-Whitney (*Analise: Nonparametric Tests: Two Independente Samples*) e obtém-se o valor de *U*. O TDE-LC calcula-se:

$$\text{TDE-LC} = \frac{U}{mn}$$

Tabela 4

Interpretação dos Tamanhos do Efeito através do U_3 de Cohen, Percentagem de Sobreposição e Probabilidade de Superioridade e do Número dos que Necessitam de Ser Tratados

Tamanho do efeito	Percentis/ U_3 de Cohen	% de sobreposição	Probabilidade de superioridade (TDE-LC)	NNT	Tamanho do efeito	Percentis/ U_3 de Cohen	% de sobreposição	Probabilidade de superioridade (TDE-LC)	NNT
0	50,0	100,0	50,0	∞	1,6	94,5	42,4	87,1	1,7
0,1	54,0	96,0	52,8	34,3	1,7	95,5	39,5	88,5	1,7
0,2	57,9	92,0	55,6	16,5	1,8	96,4	36,8	89,9	1,6
0,3	61,8	88,1	58,4	10,6	1,9	97,1	34,2	91,0	1,5
0,4	65,5	84,2	61,1	7,7	2,0	97,7	31,7	92,1	1,5
0,5	69,2	80,3	63,8	6,0	2,1	98,2	29,4	93,1	1,4
0,6	72,6	76,4	66,4	4,9	2,2	98,6	27,1	94,0	1,4
0,7	75,8	72,6	69,0	4,1	2,3	98,9	25,0	94,8	1,4
0,8	78,8	68,9	71,4	3,5	2,4	99,2	23,0	95,5	1,4
0,9	82,6	65,3	73,8	3,1	2,5	99,4	21,1	96,2	1,3
1,0	84,1	61,7	76,0	2,8	2,6	99,5	19,4	96,7	1,3
1,1	86,4	58,2	78,2	2,5	2,7	99,7	17,7	97,2	1,3
1,2	88,5	54,9	80,2	2,3	2,8	99,7	16,2	97,6	1,3
1,3	90,3	51,6	82,1	2,1	2,9	99,8	14,7	98,0	1,3
1,4	91,9	48,4	83,9	2,0	3,0	99,9	13,4	98,3	1,3
1,5	93,3	45,3	85,6	1,8	3,1	99,9	12,1	98,6	1,3

Notas: Baseado em Coe (2007), Cohen (1988), Kraemer (2006), Durlak (2009), Kraemer e Kupfer (2006), Lipsey (2012), Schünemann et al. (2008) e Wyrwich et al. (2005). Os percentis correspondem ao U_3 de Cohen. TDE-LC = tamanho do efeito em linguagem comum. NNT = número de pessoas que necessita de tratamento (assume-se que 20% do grupo de controlo tenha "resultados favoráveis", melhorando acima ou abaixo de um determinado ponto de corte.

em que *U* é a estatística *U* de Mann-Whitney, *m* é o número de participantes na primeira amostra e *n* o número de pessoas na segunda amostra (Acion, Peterson, Temple e Arndt, 2006; Furukawa e Leucht, 2011).

Equivalência percentílica e de não sobreposição. De forma semelhante, os TDE podem ser convertidos em percentis através de uma tabela de áreas sob a curva normal. A média de uma amostra de controlo com distribuição normal está no percentil 50 com uma pontuação *Z* de zero, depois adiciona-se o TDE à pontuação *Z*, segue-se a sua conversão no percentil equivalente através da tabela de áreas sob a curva normal, obtendo-se assim o índice U3 de Cohen (Cohen, 1988; Lipsey et al., 2012). Na Tabela 4 podem observar-se as percentagens de casos do grupo experimental ou de intervenção acima da média do grupo de controlo, assumindo-se uma distribuição normal. Assim, por exemplo, um TDE de 1,20 indica que a média do grupo alvo se situa no percentil 88 do grupo de comparação. Se não existir nenhum efeito, então, quer o grupo alvo, quer o grupo de comparação situam-se no percentil 50; isto é, a média do grupo alvo situa-se no percentil 50 do grupo de comparação (Durlak, 2009). Note-se que não se devem interpretar os TDE em termos de percentis quando o pressuposto da distribuição normal for infringido (Coe, 2002). Os TDE podem também ser interpretados em termos de percentagem de não-sobreposição entre o grupo alvo e o grupo de comparação (Cohen, 1988). Um TDE de zero significa que a distribuição de pontuações no grupo alvo se sobrepõe totalmente à distribuição de pontuações do grupo de comparação; ou seja, há 0% de não-sobreposição. Por ser mais compreensível, recorremos ao conceito de percentagem de sobreposição (Reiser e Faraggi, 1999) que apresentamos também na Tabela 4.

Número dos que necessitam de ser tratados. Finalmente, quando se trata da comparação de um grupo de tratamento com um grupo de comparação usa-se o *number needed to treat* (Cook e Sackett, 1995) que corresponde ao número de pacientes/utentes, por exemplo, que é necessário tratar por forma a alcançar, em média, um paciente/utente com melhoria franca (Tabela 4). Ou seja, é uma forma de apresentar valores absolutos. Para alguns (Kraemer e Kupfer, 2006), este índice é o que melhor reflete a significância clínica em termos binários (sucesso/falhanço).

Este cálculo implica identificar as pessoas que alcançaram ou excederam o limiar de mudança importante (*important change threshold*), quer no grupo de experimental, quer no grupo de comparação (Kraemer e Kupfer, 2006; Schünemann et al., 2008; Wyrwich et al., 2005).

O número que necessita tratamento define-se como o número de pacientes que se espera tratar com T (S_T) para ter mais sucesso (e menos falhanço) do que o número tratado com C (S_C), onde NNT = S_T - S_C. Para um T melhor que C, então o NNT varia do valor ideal de 1 até ao infinito; para o T pior do que C, então o NNT varia entre -1 a menos infinito (Kraemer e Kupfer, 2006, p. 992). A fórmula de conversão do d de Cohen em NNT é dada por Furukawa e Leucht (2011):

$$\text{NNT} = \frac{1}{\Phi\,[\delta - \Psi(\text{TEC}) - \text{TEC}]} \text{ ou } \text{NNT} = \frac{1}{\text{TEE} - \text{TEC}}$$

Nesta fórmula, o Φ é a função de distribuição cumulativa da distribuição normal padronizada e Ψ é a sua inversa, o TEE é a taxa de eventos no grupo experimental, o TEC é a taxa de eventos no grupo de controlo e o δ é o d de Cohen populacional. Na Tabela 4, o TEC está definido para 20%. Note-se que o NNT funciona melhor com variáveis de resultado dicotómicas (Ferguson, 2009).

Exemplo. Paiva e colaboradores (2013), entre outras análises, verificaram a influência do sexo nos comportamentos de risco e auto-dano em adolescentes através do *Risk-Taking and Self-Harm Inventory for Adolescents* (pontuação máxima = 36 pontos). O estudo mostrou que os rapazes (n_1 = 152) tinham significativamente (t = 4,61; p < 0,001) mais comportamentos de risco (M = 7,11; DP = 6,36) do que as raparigas (n_2 = 194; M = 4,47; DP = 4,26). Comecemos pela diferença não-padronizada das médias: M_1 - M_2 = 2,64. Como o desvio-padrão é grosseiramente semelhante, e os grupos têm também dimensão similar, pode calcular-se o *d* de Cohen. Determinemos então primeiro o desvio-padrão combinado:

$$DP_{\text{combinado}} = \sqrt{\frac{(152-1)6{,}36^2+(194-1)4{,}26^2}{152+194-2}} = \sqrt{\frac{9650{,}826}{344}} = 5{,}2967$$

De seguida, usemos o desvio-padrão combinado na fórmula do *d* de Cohen:

$$d \text{ de Cohen} = \frac{7{,}11-4{,}47}{5{,}30} = 0{,}498$$

O intervalo de confiança a 95% para este TDE é de [0,28, 0,71]. Com base nos critérios de Cohen (Cohen, 1988, p. 26), considerar-se-ia este TDE como *moderado*. O TDE em linguagem comum indica que, depois de controlar as diferenças individuais, a probabilidade de um rapaz ter uma média maior do que uma rapariga é de 63,8%. Quanto à interpretação percentílica, 69,2% dos rapazes estão acima da média do grupo das raparigas. Finalmente, a distribuição de pontuações no grupo dos rapazes apresenta uma percentagem de sobreposição de 80,3% em relação à distribuição de pontuações do grupo das raparigas (vide Folha de Cálculo e Tabela 4).

CAUTELAS A TER NO USO DOS TAMANHOS DO EFEITO

A interpretação e a categorização dos TDE, como referimos atrás, não deve ser rígida, pois os seus valores e as métricas usadas são arbitrários (Andersen, McCullagh e Wilson, 2007; Blanton e Jaccard, 2006a; 2006b; Embretson, 2006; Kazdin, 2006; Thompson, 2007) e há que ter em consideração as áreas de investigação (Lipsey et al., 2012; McCartney e Rosenthal, 2000; Zakzanis, 2001), e o contexto de variáveis da Realidade (Aguinis, 2010; Lipsey et al., 2012). O TDE, há que recordá-lo, é simplesmente uma transformação da diferença entre médias em unidades do desvio-padrão.

Arbitrariedade das medições

O problema nas ciências sociais e comportamentais, e especialmente em psicologia, é não existirem unidades para constructos abstratos e não-físicos associadas ao comportamento no mundo real (p. e., auto-conceito, inteligência, sintomas depressivos, ou satisfação com a vida), nem existirem medidas universalmente aceites para esses constructos (Andersen et al., 2007; Kline, 2004). Por isso, este tipo de métricas são consideradas *arbitrárias*, isto é, avaliadoras indiretas de constructos hipotéticos não observados (Andersen et al., 2007; Blanton e Jaccard, 2006a; Cohen, 1988; Kline, 2004).

Este problema da arbitrariedade das medidas é especialmente evidente nos estudos de intervenção (Kazdin, 2006). Em 1996, Sechrest, McKnight e McKnight tinham já discutido este aspeto, exemplificando com a intervenção na depressão. Neste tipo de estudo, medem-se, por exemplo, os sintomas depressivos com o *Beck Depression Inventory* (BDI) antes da intervenção, depois colocam-se os participantes com sintomas nos grupos de tratamento e de comparação, e de seguida usa-se o BDI novamente após o tratamento. Se o grupo de tratamento tiver uma pontuação inferior à do grupo de comparação no final do tratamento, então a terapia seria considerada eficaz. Em termos metodológicos, pouco há a apontar. Mas o que significa a descida no BDI? Para o doente, uma melhoria real, não será uma descida nas pontuações de um qualquer instrumento, mas sim, por exemplo, conseguir trabalhar, passar mais tempo ativo, andar mais alegre, ter mais interesse em conviver e dormir melhor.

Para maior aprofundamento destas questões sobre a arbitrariedade das medições através de testes psicométricos há uma vasta literatura ao dispor (Andersen, McCullagh e Wilson, 2007; Blanton e Jaccard, 2006a; 2006b; Embretson, 2006; Kazdin, 2006).

Necessidade de Interpretar os Tamanhos do Efeito no Contexto da Área de Investigação

Para Zakzanis (2001), só se devem usar as classificações de *pequeno*, *moderado*, ou *grande* quando a área do estudo se compreende em termos de TDE. Para outros autores, quando se procura interpretar se os TDE são *pequenos*, *moderados*, ou *grandes*, há que fazê-lo contrastando com as normas apropriadas; baseando-se essas normas nas distribuições dos TDE obtidas com medições comparáveis, em intervenções comparáveis dirigidas a amostras também comparáveis (Lipsey et al., 2012).

Em síntese, a interpretação aprofundada dos TDE deve ter em consideração o *contexto* em que os valores foram obtidos (p. e., instituição, escola, clínica), o *planeamento* do estudo (p. e., estudo experimental), escolhas *metodológicas* (p. e., qual a variável preditora), o quadro da *métrica* usada (p. e., quociente de inteligência), os outros efeitos na *literatura* (p. e., diferença padronizada da média maior), os efeitos de intervenções *semelhantes* (p. e., 20% mais eficiente) ou outros TDE que sejam bem compreendidos conceptualmente (Cohen, 1988; McCartney e Rosenthal, 2000; Borenstein, 2009, p. 234; Lakens, 2013). Por exemplo, em estudos de intervenção em educação é raro haver TDE acima dos 0,30 (Lipsey et al., 2012). Já Zakzanis (2001, p. 664) propõe um valor de *d* acima de 3,00 (efeito "grande") para discriminar um grupo experimental de um grupo de controlo no domínio da neuropsicologia, recomendando que efeitos entre 0,1 e 2,0 devem ser interpretados à luz do contexto.

Apesar da importância da comparação com os valores obtidos noutros estudos, executando, por exemplo uma meta-análise, a comparação deve seguir diversas cautelas (p. e., fidedignidade das medições; homogeneidade populacional; ou "força do tratamento"; Olejnik e Algina, 2000). Há que ter ainda em atenção que, quando se faz a revisão da literatura, há uma tendência para se encontrar somente estudos em que houve resultados *positivos*, isto é, estudos com resultados estatisticamente significativos. Aquilo a que Rosenthal (1979) designou pelo "file drawer problem": os estudos com resultados que suportam a hipótese nula vão parar à gaveta do arquivo dos ficheiros mortos. Para além de que muitos autores têm dificuldades em publicar estudos com resultados estatisticamente não significativos (Wolf, 1986).

Orwin (1983) forneceu uma fórmula de segurança (N_{fs}) que permite ultrapassar parcialmente esta questão, ao calcular o número de estudos confirmadores da hipótese nula que seriam necessários para reverter a conclusão de que existe uma relação significativa:

$$N_{fs} = \frac{N\,(d - d_c)}{d_c}$$

sendo N o número de estudos na meta-análise, d o TDE médio determinado a partir dos estudos em análise e d_c o valor de critério selecionado de forma a igualar o d quando um número conhecido de estudos hipotéticos (N_{fs}) fosse adicionado à meta-análise. Orwin (1983) sugeriu que o valor de critério fosse 0,2 [seguindo a sugestão de Cohen (1988) para um efeito pequeno]. Assim, N_{fs} calcula o número necessário de estudos hipotéticos para ultrapassar um valor de 0,2.

Necessidade de Interpretar os Tamanhos do Efeito no Contexto de Variáveis da Realidade

Quando se apresentam os aspetos quantitativos dos TDE, há que depois perceber o significado prático desses valores através de uma análise qualitativa que descreva a importância dos resultados no seu contexto (Aguinis, 2010; Lipsey et al., 2012). Assim, um pequeno TDE pode ter consequências importantes, como, por exemplo, o de uma intervenção que diminua a taxa de suicídio (Lakens, 2013).

Repare-se no exemplo que apresentámos na secção anterior sobre a influência do sexo nos comportamentos de risco e auto-dano em adolescentes. O que significa a classificação de *moderado* (d = 0,498) numa métrica que não tem um significado direto? Este valor, como é habitual quando se usam questionários, não é, nem diretamente interpretável, nem significativo na *prática* (Kirk, 1996). A classificação também não nos responde à questão de saber se a diferença é relevante *clinicamente* (Jacobson e Truax, 1991)[xvi].

Tome-se agora outro exemplo de intervenção, mas em neuropsicologia: reuniu-se uma amostra de idosos com declínio cognitivo sem demência e aplicou-se a metade desses idosos um programa de reabilitação neuropsicológica (a outra metade ficou em lista de espera). Calculou-se o TDE nas pontuações de uma escala depressiva entre o grupo de intervenção e o grupo de comparação e obteve-se um valor de 0,20 (Lemos et al., 2014). Este pequeno valor significa que 58% dos idosos do grupo de intervenção ficaram acima da média dos idosos do grupo de controlo. Quanto ao NNT, para se ter mais um resultado favorável no grupo de intervenção em contraste com o grupo de comparação, seria preciso integrar no programa de reabilitação 16,5 pessoas. O que significa que se 100 pessoas forem integradas no programa de tratamento, 6,1 mais pessoas tenderão a ter um resultado favorável comparado com o que teriam se não tivessem reabilitação neuropsicológic a (cf. Tabela 4)[xvii]. Quando comparamos o TDE obtido com o de um estudo de intervenção equivalente (0,04; Liesbeth et al.,

2011), percebemos que a eficácia do programa português foi maior. Podemos especular porquê. Terá sido o tipo de intervenção? A duração da reabilitação? Apesar de não se ficar a perceber qual o real impacto na vida do dia a dia dos participantes no estudo, se nos cingirmos somente à significância estatística, não poderíamos então comparar com outros estudos, e teríamos menos dados para equacionar novas hipóteses a testar.

CONCLUSÕES

Devido às limitações dos testes estatísticos e do valor de *p* para informar sobre o real significado dos resultados de investigação, desde há muito que se defende que se apresentem os tamanhos de efeito. Vimos que é possível existir uma diferença "não-significativa", sem que isso seja o mesmo que ter um "não-efeito". Descrevemos também como é rara a apresentação destes efeitos. Provavelmente isso deve-se ao facto de poucos investigadores terem uma ideia clara de como proceder ou interpretá-los Zakzanis (2001). Neste artigo, não só argumentamos a importância de apresentar as magnitudes do efeito no cálculo das diferenças, como mostrámos também as fórmulas para a sua computação acompanhadas por uma folha de cálculo em excel para agilizar o processo. Vimos ainda como interpretar estes tamanhos em termos de *linguagem comum* e em termos percentílicos e, ainda, o número dos que necessitam de ser tratados. Defendemos também a importância de interpretar as magnitudes do efeito no contexto do estudo, em vez de recorrer imediatamente às classificações propostas por Cohen, e de procurar extrapolar o seu significado para a realidade.

Em síntese, apresentar e interpretar TDEs não deve ser visto como mais uma barreira a ultrapassar no mundo editorial (Durlak, 2009). Ainda assim, as autoras deste artigo esperam que os autores da RPICS passem a incluir e a interpretar os TDE nas investigações em que tal seja apropriado e que este artigo forneça as diretrizes necessárias e suficientes para esse propósito. A súmula dessas diretrizes é apresentada no Quadro 1.

Quadro 1
Diretrizes para Apresentar e Interpretar os Tamanhos do Efeito (TDE)

1. Escolher o TDE mais adequado ao tamanho da amostra de cada grupo
2. Indicar os dados descritivos básicos (média e desvio-padrão)
3. Calcular o TDE, independentemente dos resultados serem ou não significativos
4. Designar o tipo de TDE usado, indicar as referências apropriadas e/ou apresentar a equação
5. Calcular o Intervalo de confiança a 95% para o TDE
6. Classificar o TDE de acordo com os critérios de Cohen (1988) e de Rosenthal (1996)
7. Interpretar o TDE em linguagem comum (uma das seguintes hipóteses)
   a. Converter em percentis
   b. Calcular a percentagem de sobreposição
   c. Calcular a probabilidade de superioridade
8. Interpretar o TDE no contexto da área de investigação
   a. Rever a literatura, optando por estudos com planeamento e métodos de cálculo do TDE semelhantes
   b. Calcular os TDEs dos estudos revistos quando ausentes
9. Interpretar o TDE no contexto de variáveis da Realidade

REFERÊNCIAS

Acion, L., Peterson, J. J., Temple, S. e Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25(4), 591–602. doi:10.1002/sim.2256

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H. e Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*(3), 515–539.

Aickin, M. (2004). Bayes without priors. *Journal of Clinical Epidemiology*, 57(1), 4–13. doi:10.1016/S0895-4356(03)00251-8

American Psychological Association (APA) (2010). *Publication Manual of the American Psychological Association* (6.ª ed.). Washington, DC: APA.

Andersen, M. B., McCullagh, P. e Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport e Exercise Psychology*, 29(5), 664–672.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617.

Berben, L., Sereika, S. M. e Engberg, S. (2012). Effect size estimation: methods and examples. *International Journal of Nursing Studies*, *49*(8), 1039–1047. doi:10.1016/j.ijnurstu.2012.01.015

Bezeau, S. e Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section a)*, *23*(3), 399–406.

Blanton, H. e Jaccard, J. (2006a). Arbitrary metrics in psychology. *The American Psychologist*, *61*(1), 27–41. doi:10.1037/0003-066X.61.1.27

Blanton, H. e Jaccard, J. (2006b). Arbitrary metrics redux. *The American Psychologist*, *61*(1), 62.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hodges e J. C. Valentine, *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York: Russell Sage Foundation.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management*, *29*(1), 79–97.

Caperos, J. M. e Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema*, *25*(3), 408–414. doi:10.7334/psicothema2012.207

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*(1), 105-110.

Coe, R. (2002). *It's the effect size, stupid: what effect size is and why it is important.* Presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, Education-line.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2.ª ed.). Hillsdale: Lawrence Erlbaum Associates.

Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*(1), 155.

Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101.

Cohen, J. (1994). The earth is round (p < .05). *The American Psychologist*, *49*(12), 997-1003.

Conn, V. S., Chan, K. C. e Cooper, P. S. (2014). The problem with p. *Western Journal of Nursing Research*, *36*(3), 291–293.

Cook, R. J. e Sackett, D. L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *BMJ*, *310*(6977), 452–454.

Cooper, H., Hedges, L. V. e Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2.ª ed.). New York: Russell Sage Foundation.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin*, *116*(3), 509–511.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928.

Ellis, P. D. (2010). *The essential guide to effect sizes. Statistical power, meta-analysis, and the interpretation of research results* (pp. 1–193). Cambridge: Cambridge University Press.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *The American Psychologist*, *61*(1), 50–55. doi:10.1037/0003-066X.61.1.50

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532-538.

Fern, E. F. e Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*(2), 89–105.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1959). *Statistical methods and scientific inferente* (2.ª ed.). Edinburgh: Oliver and Boyd.

Furukawa, T. A. e Leucht, S. (2011). How to obtain NNT from Cohen's d: Comparison of two methods. *PLoS ONE*, *6*(4), e19070, 1-5.

Giere, R. N. (1972). The significance test controversy. *British Journal for the Philosophy of Science*, *23*(2), 170–181.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Glass, G.V., McGaw, B. e Smith, M. L. (1981). *Meta-analysis in social research.* Sage: Beverly Hills.

Grissom, R. J. e Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128.

Hedges, L. V. e Olkin, I. (1985). *Statistical methods for meta-analysis* (Vol. 11, pp. 104–106). Orlando: Academic Press.

Hentschke, H. e Stüttgen, M. C. (2011). Computation of measures of effect size for neuroscience data sets. *The European Journal of Neuroscience*, *34*(12), 1887–1894. doi:10.1111/j.1460-9568.2011.07902.x

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education*, *61*(4), 317–333.

Jacobson, N. S. e Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12-19.

Kazdin, A. E. (2006). Arbitrary metrics: implications for identifying evidence-based treatments. *The American Psychologist*, *61*(1), 42-49. doi:10.1037/0003-066X.61.1.42

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345–353.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759. doi:10.1177/0013164496056005002

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research* (2.ª ed.). Washington, DC: American Psychological Association.

Kraemer, H. C. e Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *BPS*, *59*(11), 990–996. doi:10.1016/j.biopsych.2005.09.014

Kühberger, A., Fritz, A. e Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), e105825, 1-8. doi:10.1371/journal.pone.0105825

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1-12. doi:10.3389/fpsyg.2013.00863

Lee, M. D. e Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668. doi:10.1037/0033-295X.112.3.662

Lemos, L., Espirito-Santo, H., Silva, G. F., Costa, M., Cardoso, D., Vicente, F., et al. (2014). *The impact of a Neuropsychological Rehabilitation Group Program (NRGP) on cognitive and emotional functioning in institutionalized elderly* (p. 1). Presented at the 22nd European Congress of Psychiatry, Munich.

Lenth, R. V. (2006–2014). *Java applets for power and sample size.* Acedido em http://homepage.stat.uiowa.edu/~rlenth/Power/

Liesbeth, W. A., Prins, J. B., Vernooij-Dassen, M. J. F. J., Wijnen, H. H., Olde Rikkert, M. G. M. e Kessels, R. P. C. (2011). Group therapy for patients with mild cognitive impairment and their significant others: results of a waiting-list controlled trial. *Gerontology*, *57*(5), 444–454. doi:10.1159/000315933

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W. … Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*. National Center for Special Education Research, Institute of Education Sciences.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, *36*(2), 102–105.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 161–171.

McCartney, K. e Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, *71*(1), 173–180.

McGraw, K. O. e Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361-365.

McMillan, J. H. e Foley, J. (2011). Reporting and discussing effect size: Still the road less traveled. *Practical Assessment, Research e Evaluation*, *16*(14), 1–12.

Morrison, D. E. e Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader.* Chicago: Aldine.

Nakagawa, S. e Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. doi:10.1111/j.1469-185X.2007.00027.x

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. doi:10.1037//1082-989X.5.2.241

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Olejnik, S. e Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286. doi:10.1006/ceps.2000.1040

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

Paiva, A. C., Cunha, M., Xavier, A. M., Marques, M., Simões, S. e Espirito-Santo, H. (2013). Exploratory study of risk-taking and self-harm behaviours in adolescents: prevalence, characteristics and its relationship to attachment styles. *European Psychiatry*, *28*(Supl. 1). doi:10.1016/S0924-9338(13)76530-1

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series a, Containing Papers of a Mathematical or Physical Character*, *195*, 1–47. doi:10.1098/rsta.1900.0022

Reiser, B. e Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal equal variance case. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *48*(3), 413–418.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Rosenthal, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology, 51*, 4-13.

Rosenthal, R. (1994). Parametric measures of effect size. Em H. Cooper e L. V. Hedges (Eds.). *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37-59.

Rosnow, R. L. e Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *The American Psychologist*, *44*(10), 1276-1284.

Rosnow, R. L., Rosenthal, R. e Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*(6), 446–453.

Salsburg, D. (2002). *The lady tasting tea*. New York: Macmillan.

Sanabria, F. e Killeen, P. R. (2007). Better statistics for better decisions: rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in the Schools*, *44*(5), 471–481. doi:10.1002/pits.20239

Schatz, P., Jay, K. A., McComb, J. e McLaughlin, J. R. (2005). Misuse of statistical tests in Archives of Clinical Neuropsychology publications. *Archives of Clinical Neuropsychology*, *20*(8), 1053–1059. doi:10.1016/j.acn.2005.06.006

Schmidt, F. L. e Hunter, J. E. (2004). *Methods of Meta-Analysis*. Thousand Oaks: SAGE Publications.

Schneider, A. L. e Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. *Evaluation Review*, *8*(4), 573–582. doi:10.1177/0193841X8400800407

Schünemann, H. J., Oxman, A. D., Vist, G. E., Higgins, J. P. T., Deeks, J. J., Glasziou, P. e Guyatt, G. H. (2008). Interpreting results and drawing conclusions. Em J. P. T. Higgins e S. Green, *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series* (pp. 1–29). The Cochrane Collaboration.

Sechrest, L. McKnight, P. e McKnight, K. (1996). Calibration of measures in psychotherapy outcome studies. *American Psychologist, 51,* 1065-1071.

Sedlmeier, P. e Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309-316.

Snyder, P. e Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, *61*(4), 334–349.

Sun, S., Pan, W. e Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989-1004.

Tabachnick, B. G. e Fidell, L. S. (2007). *Using Multivariate Statistics* (5.ª ed.). Boston: Pearson.

NOTAS

i Um valor de *p* de 0,05 significa que as chances de obter uma diferença média dessa magnitude ou maior é de 1 em 20 e as chances de obter uma diferença média dessa magnitude ou menor é de 19 em 20 se as duas amostras representarem a mesma população (Carver, 1978, p. 392).

ii Entre as alternativas ao nível de significância, queremos destacar a sugestão de estimar a probabilidade de replicação (Sanabria e Killeen, 2007) e de replicar efetivamente o estudo com outras amostras (Carver, 1978). Outra das alternativas propõe a aplicação do paradigma Bayesiano Lee e Wagenmakers, 2005; Wagenmakers, 2007).

iii "… Deus gosta do 0,06 quase tanto como do 0,05".

iv "não há uma razão lógica forte por detrás desta opção".

v Nas palavras de Conn (2014, p. 291): força a tomada de decisões binárias (tipo sim-não) para dados contínuos.

vi Para aprofundar a temática do poder dos testes, o leitor interessado pode recorrer aos textos de Cohen (1988, pp. 14-19; 1992b, p. 100) e de Kline (2004, pp. 76-81).

vii Qual é a probabilidade de obter uma pessoa morta (etiquete esta parte D) dado que a pessoa foi enforcada (rotule esta parte H); isto é, na forma de símbolo, o que é *p*(D|H)? Obviamente, que vai ser muito alta, talvez 0,97 ou superior. Agora, vamos inverter a pergunta. Qual é a probabilidade de uma pessoa ter sido enforcada (H) dado que a pessoa está morta (D); isto é, o que é *p*(H|D)? Desta vez, a probabilidade será, sem dúvida, muito baixa, talvez 0,01 ou inferior.

viii Ainda que se deva ter cuidado com amostras pequenas, pois há uma tendência para obter tamanhos do efeito maiores com amostras pequenas (Kühberger, Fritz e Scherndl, 2014). Por isso, se sugere que se faça a análise sistemática dos TDE em diferentes estudos por forma a determinar até que ponto o tamanho da amostra influencia os resultados (Snyder, 1993).

ix "apresente sempre o tamanho do efeito para resultados primários".

x "a notificação e interpretação dos TDE [tamanhos do efeito] no contexto dos efeitos previamente relatados é essencial para uma boa pesquisa".

xi Para reduzir potenciais efeitos de enviesamento de artefactos estatísticos e aproximar o TDE, baseado numa amostra, a um TDE populacional, vejam-se as sugestões de Breaugh (2003, p. 94-95) ou, de forma mais detalhada, as de Schmidt e Hunter (2004).

xii O *tamanho da amostra* depende do *nível de significância* (pode ser α = 0,05), do erro padrão (para isso obtêm-se os desvios-padrão obtidos por outros estudos que usaram a mesma medição e depois divide-se o desvio-padrão pelo *n* para obter o erro padrão) e do *poder* do teste (sugere-se o valor de 0,90 ou mesmo de 0,95) e do tamanho da população (Cohen, 1992a). Para um poder de 80%, são necessários 393 sujeitos para um efeito pequeno, 64 para um médio e 26 para um grande.

xiii "há um certo risco inerente em oferecer definições operacionais convencionais para esses termos a usar na análise do poder num campo de investigação tão diverso como é a ciência comportamental".

xiv "algum sentido à sua escala [do TDE]"

xv "um índice atrativo do tamanho do efeito que não requer um conhecimento prévio de estatística para o entender".

xvi De acordo com a revisão de Jacobson (1991, p. 12), no domínio da intervenção psicoterapêutica, os critérios para a significância clínica devem incluir: uma grande percentagem de melhoria; um grau de mudança reconhecível pelas pessoas significativas; a eliminação do problema; níveis normativos de funcionamento; ou mudanças que reduzam o risco de outros problemas de saúde.

xvii Assume-se que 20% do grupo de comparação tenha "resultados favoráveis", melhorando acima de um ponto de corte definido.