Psicothema

Psicothema 2020, Vol. 32, No. 3, 410-419
doi: 10.7334/psicothema2019.396

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Change in the center of the distribution and in the individual scores: Relation with heteroskedastic pre- and post-test distributions

Eduardo Estrada[1], José Manuel Caperos[2,3], and Antonio Pardo[1]
[1] Universidad Autónoma de Madrid and [2] Fundación San Juan de Dios, and [3] Universidad Pontificia de Comillas

Abstract

Background: Although average-based effect size (ES) and percentage of individual changes (PIC) are quite different, they are not independent: larger ESs lead to higher PICs. However, this association has not been sufficiently explored. **Method:** We analyzed this association based on data simulated in the context of a pre-post design, with and without control groups. We simulated various distributions, sample sizes, degrees of test-retest reliability, effect sizes, and different variances in pre- and post-test. **Results:** The PIC is closely associated with the ES across a wide variety of empirically frequent scenarios. In the "single group pre-post designs", the linear regression model shows R2 values above 0.90. In the "control group pre-post designs", the linear regression model shows R2 values above 0.80. These results were found even when the post-test variability differed from that of the pre-test, replicating, extending and generalizing the findings in previous studies. **Conclusions:** (1) In the absence of information about the PIC, the ES may be used to estimate this percentage. (2) The PIC is useful in interpreting the meaning of ES measures.

Keywords: Individual change, group change, effect size, percentage of changes, reliable change index.

Resumen

Cambio en el centro de la distribución y en las puntuaciones individuales: relación con distribuciones heteroscedasticas pre y post prueba. **Antecedentes:** aunque el tamaño del efecto (ES) y el porcentaje de cambios individuales (PIC) son cosas distintas, no parecen ser independientes: mayores ESs conllevan mayores PICs. Pero esta relación todavía no ha sido suficientemente explorada. **Método:** estudiamos dicha relación mediante datos simulados en el contexto de un diseño pre-post con y sin grupo control. En la simulación se han utilizado diferentes distribuciones, tamaños muestrales, niveles de fiabilidad test-retest, efectos de varios tamaños y distintas variabilidades en el pre- y en el post-test. **Resultados:** el PIC está estrechamente relacionado con el ES. En los diseños pre-post, el modelo de regresión lineal ofrece valores R2 por encima de 0,90. En los diseños pre-post con grupo control, valores R2 por encima de 0,80. Estos resultados se mantienen incluso cuando la variabilidad del post-test es distinta de la del pre-test. **Conclusiones:** (1) cuando no se tiene información sobre el PIC, el tamaño del efecto puede utilizarse para estimar ese porcentaje; (2) el PIC sirve para precisar el significado de las medidas del tamaño del efecto.

Palabras clave: cambio individual, cambio grupal, tamaño del efecto, porcentaje de cambios, índices de cambio fiable.

In various fields of psychology, and especially in the clinical setting, it is common to perform interventions. Evaluating whether interventions have an effect generally requires determining when a change occurs. However, there is no single method to establish the occurrence of a change (see, for example, Ogles, Lunnen, & Bonesteel, 2001). In addition to the therapist's criteria, different sources can be taken into consideration.

It is very common to study change through the analysis of the patients' answers to scales or questionnaires (*distribution-based methods*, see Crosby, Kolotkin, & Williams, 2003). This is a widely-used strategy in psychology, with a long tradition of using tests to obtain measurements. It has gained momentum over the last two decades also, among others, in the field of medicine to evaluate health-related quality of life (Norman, Sridhar, Guyatt, & Walter, 2001).

However, at least two perspectives can be adopted for assessing whether the answers to a scale or questionnaire show a reliable change: the *group* or the *individual*. Both perspectives are applicable to pre-post designs. In the *group* perspective, the goal is to evaluate whether the group, as a whole, has experimented a change. This is usually done by comparing the pre- and post-test means (or other measures of central tendency) using a significance test or an effect size (ES) measure (Cohen, 1988; Grissom & Kim, 2012; Pek & Flora, 2018). This strategy is based on the *center of the distributions*, and changes detected through such strategy has been termed *average based change* (Estrada, Ferrer, & Pardo, 2019).

The *individual* perspective evaluates which particular individual experienced a change by applying indices that specify the smallest change that cannot be attributed to random sample fluctuations or measurement errors (Crosby et al., 2003; Jacobson & Truax, 1991). This minimal change amount is usually referred to as the *statistically reliable change, minimal detectable change*, or, simply, *reliable change* (see, for example, Beaton, Bombardier,

Received: December 12, 2019 • Accepted: February 2, 2020
Corresponding author: Eduardo Estrada
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: eduardo.estrada@uam.es

Katz, & Wright, 2001; de Vet et al., 2006; Osma, Sánchez-Gómez, & Peris-Baquero, 2018). This strategy, based on *individual scores*, has been termed *individual based change* (Estrada et al., 2019)

These two perspectives appear to provide quite different information. For example, Schmitt and Di Fabio (2004, pp. 1008-1009), claimed that "statistically significant change at the group level may not be significant at the individual level (…). Mean changes for a group may be the result of few individuals with relatively large changes, or numerous individuals with relatively small changes". Similarly, Vindras, Desmurget and Baraduc (2012, p. 2) mentioned that "the effect of a factor can be significant for every individual (compared to intra-individual variability) while Student and Fisher tests yield a probability close to one if the population average is small enough".

These previous works illustrate the assumption (frequent in clinical literature) that group and individual change are different, because the change in the center of the distribution does not inform about which particular individuals changed. However, there is sound evidence that a strong association exists between the two approaches: larger displacements from the center of the distribution are associated with a larger percentages of individual changes (PIC).

For example, Norman et al. (2001) found that the effect size following an intervention (i.e., the amount of change in the center of the distribution) is the factor with the strongest association with the percentage of subjects who benefit from the intervention. Lemieux et al. (2007) came to a similar conclusion, using real data, instead of simulated.

In a recent and broader simulation study, Estrada et al. (2019), found a strong association between group change (*ES*, quantified as *effect size measurements*) and the percentage of individual changes (*PIC*, computed from *reliable change indices*). These results allowed specifying: (a) the function describing the association between the two perspectives; and (b) the fit of such function.

Although Estrada et al. (2019) found a strong association in a wide variety of realistic conditions, they did not explore situations involving a pre-post change in the scores' variability. They modified the distributions' centers, but kept their variances constant across time points. However, it is entirely possible (and frequent in empirical scenarios) to find a different degree of dispersion in the pre- and post-test distributions. Indeed, the scores' variability can change as an effect of the intervention applied.

For example, Foster, Harrison, Draheim, Redick, and Engle (2017) found a "magnification effect" in a working memory training study: participants with higher initial levels showed larger gains after a 20-session program. In contrast, other researchers have proposed a "compensation effect": Training strategies have a greater impact on performance when subjects' baseline performance is low (for a succinct review, see Smoleń, Jastrzebski, Estrada, & Chuderski, 2018). Similarly, in the clinical field, it is reasonable to expect that, when treating individuals who suffer depression, those with less acute levels will show smaller gains because they are closer to the nonclinical population. These individual differences in pre-post change can lead, in turn, to an increase (magnification) or decrease (compensation; Macías, Valero-Aguayo, Bond, & Blanca, 2019) in the distribution's variability. The available literature provides no information about the relation between average and individual based change in these relevant scenarios.

Therefore, the present study pursues to extend what is already known by: (a) evaluating whether the association found in previous studies between average based change (ES) and individual based change (PIC) is replicated when the pre- and post-test distributions have different variability; (b) if the association is replicated, describing its shape, proposing a mathematical function able to capture it, and quantifying the fit of such function; and (c) determining under which conditions the association can be found (normality, slight, moderate or severe departures from normality, etc.) and how different conditions affect it (in particular, the change variability from pre- to post-test).

Method

Procedure

We simulated data corresponding to two of the most commonly-used designs in the health and behavioral sciences field: a "single group pre-post design" (*pre-post* design) and a "pre-post design with control group" (*control-pre-post* design). Therefore, we generated scenarios in which a group of subjects (two groups in control-pre-post design) had the same variable measured at two different time points (generally, before and after an intervention) with the goal of evaluating whether a change occurred between them. Including a single group design is important because it is common in applied contexts, and the indices for individual change were developed in this context. On the other hand, including a control group (ideally randomly assigned) is the only way to attribute change to the intervention (Shadish, Cook, & Campbell, 2002).

Simulated conditions. We manipulated the following criteria:

a. *Change in the center of the distribution* (i.e., *effect size*). Quantified using the standardized mean of the pre-post differences: *d* (see the section *Data Analysis* below). It ranged between 0 and 3.6, in steps of 0.3 points. These values were chosen to enable gathering information on a wide range of effects: from a null effect to an extremely large one (allowing the PIC to range between 0 and 100%). With the exception of the "null effect" condition, we assumed that the average scores increased between pre- and post-test. Therefore, given that the pre-post differences were calculated by subtracting the pre- from the post-test score, we worked with *positive* effects. Consequently, we applied one-tailed right tests in all the conditions with non-null effects, and two-tailed tests for the condition with d=0.
   In control-pre-post designs, the effect size for the experimental group was modified according to the simulation scheme, but was always constrained to zero in the control group (see the *Simulation Process* below).
   It is important to point out that *d* was fixed for the population. This implies that: (a) the empirical values of *d* differed in each replication and (b) centered on each of the values chosen for *d*, a random distribution of individual changes was generated. Therefore, in each sample, each individual case experimented a different amount of change. The variance of that change distribution was established depending on the criteria *b* and *e* (see below).
b. *Degree of dispersion of the distribution of the experimental group post-test scores* (i.e., *standard deviation* of post-test scores of the experimental group). The pre-test scores of the experimental group were generated by assigning the value

of 1 to the standard deviation; the post-test scores for the experimental group were generated assuming a change in the variability of the scores after the intervention, with the following standard deviations: 0.25, 0.5, 1, 2 and 4. In the control group and in the pre-test of the experimental group, the standard deviation was always one.

c. *Sample size* (*n*). Three sample sizes (20, 50, 100) were chosen to examine samples that are typically considered small, medium and large (see, Crawford & Howell, 1998). In the control-pre-post design, we used groups of the same size.

d. *Pre-post correlation* (R_{XY}): 0.3, 0.5, 0.7 and 0.9. These values were chosen to represent the range of values that are usually found in applied contexts (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). The Pearson correlation coefficient was used to quantify this association. In the control-pre-post design, we imposed the same pre-post correlation in both groups.

e. *Shape of the pre- and post-test distributions*. Previous literature has shown that empirical data sets often depart from normality (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013). Therefore, we established seven values for skewness (from extreme negative, g_1=–3, to extreme positive, g_1=3) and four degrees of kurtosis (from normal, g_2=0, to extreme, g_2=18), combined as follows: (1) extreme negative skewness: g_1=–3, g_2=18; (2) moderate negative skewness: g_1=–2, g_2=9; (3) slight negative skewness: g_1=–1, g_2=2; (4) normality: g_1=0, g_2=0; (5) slight positive skewness: g_1=1, g_2=2; (6) moderate positive skewness: g_1=2, g_2=9; and (7) extreme positive skewness: g_1=3, g_2=18 (note that the degree of kurtosis is partially conditioned by the degree of skewness). Less than 5% of empirical distributions are more extreme than those simulated here (Blanca et al., 2013). In the control-pre-post design, the same shape of the distribution was imposed for both groups.

Simulation process. The combination of the five criteria described above resulted in $13 \times 5 \times 3 \times 4 \times 7$=5,460 conditions. For each of these conditions, 200 samples were generated, including one experimental group and one control group (1,092,000 samples in total). Simulation was implemented using the "MatLab 2011*a*" software.

For each replication, we generated two independent matrices, $\mathbf{X_1}=(X_1^*,Y_1^*)$ and $\mathbf{X_2}=(X_2^*,Y_2^*)$, each matrix with *n* pairs of scores in two uncorrelated variables. These scores were generated using the Pearson distribution system. Both variables were assigned the same mean, standard deviation, skewness, and kurtosis. Initially, all the population means equaled zero and the standard deviations equaled one. Skewness and kurtosis were modified systematically according to the g_1 and g_2 values described in the preceding section. The *X* and *Y* variables were generated randomly to guarantee that post-test scores could be the same, higher, much higher, lower, or much lower than the corresponding pre-test scores, as usually occurs in empirical scenarios. Therefore, this initial step ensured that each case experimented a different amount of change.

To impose the desired degree of pre-post correlation, in the second step of the process we multiplied the $\mathbf{X_1}$ and $\mathbf{X_2}$ matrices by the Cholesky decomposition (MatLab *cholcov* function) of the correlations matrix (**R**) corresponding to the chosen R_{XY} correlations (0.30, 0.50, 0.70, 0.90). As a result of this step, we obtained a matrix $\mathbf{M_1}$ with two variables (X_1=pre_exp; Y_1=post_exp) for the experimental group and a matrix $\mathbf{M_2}$ with two variables (X_2=pre_ctrl; Y_2=post_ctrl) for the control group, distributed according to pre-established conditions and correlating X_1 with Y_1 and X_2 with Y_2 with a value similar to that chosen for each condition.

Finally, we modified the variable Y_1 (experimental group post-test scores) to apply the expected effect. This modification implied adding, to all post-test scores, the result of multiplying the standard deviation of the pre-post differences by a value ranging from 0 to 3.6 points, with increments of 0.3. No value was added to the control group post-test scores.

Data analysis

First, after generating the 1,092,000 samples, we verified whether the characteristics of the simulated distributions corresponded with the simulation values by computing descriptive of central tendency, dispersion, shape, and pre-post correlations for every sample.

Second, to quantify the **group change** in the single group conditions, we calculated Cohen's (1988) *d* by dividing, in each sample, the difference between the post- and pre-test measurements (M_{pre}, M_{post}) by the standard deviation of the pre-post differences:

$$d_{pp} = \frac{M_{post} - M_{pre}}{S_{dif}}$$

("pp"=pre-post design). To quantify the group change in control-pre-post design, we used two different statistics: Hays's ω^2 and Cohen's *d*. The statistic ω^2 (omega-squared) associated with the effect of the interaction between the between-subject factor (the groups) and the within-subject factor (the pre- and post-test time points) enables capturing the difference between the groups by comparing the mean change observed in the experimental group with the mean change observed in the control group (Hays, 1988; Kirk, 2013). In a control-pre-post design:

$$\hat{\omega} = \frac{gl_{AB}\left(F_{AB}-1\right)}{gl_{AB}\left(F_{AB}-1\right)+n}$$

where F_{AB} is the *F* statistic associated with the interaction effect, gl_{AB} are the degrees of freedom and *n* is the total number of scores in the dataset.

We chose a version of Cohen's standardized difference proposed to quantify the interaction effect in the context of meta-analysis of *control-pre-post* designs (*cpp*, Grissom & Kim, 2012, pp. 90-92):

$$d_{cpp} = \frac{\left(M_{post\,exp} - M_{pre.exp}\right) - \left(M_{post.ctrl} - M_{pre.ctrl}\right)}{S_{pre}}$$

Where S_{pre} refers to the pooled standard deviation of the two pre-test groups:

$$S_{pre} = \sqrt{\left(S^2_{pre.exp} + S^2_{pre.ctrl}\right)/2}$$

Third, to obtain the PIC, we applied individual change indices to identify which individual score presented a reliable change.

Among the many reliable change indices available to evaluate the individual change in the pre-post designs, we chose two with excellent performance according to previous studies (Pardo & Ferrer, 2013; Ferrer & Pardo, 2014). First, the *standardized individual difference* (*SID*), i.e., the standardized score resulting from dividing each pre-post difference (D_i) by the standard deviation of the differences (S_{dif}):

$$SID = \frac{D_i}{S_{\text{dif}}}$$

This statistic was initially proposed by Payne and Jones (1957) to evaluate the *abnormality* of the discrepancy between two scores. If the distribution of the pre-post differences is normal, 95% of the *SID* are expected to be between ±1.96, and that 90% between ±1.645.

The second index is the *reliable change index (RCI)* proposed by Jacobson and collaborators (Jacobson & Truax, 1991). This is, probably, the best-known individual change index. It is based on the standard error of measurement. We applied a corrected version allowing for pre- and post- homoskedasticity (Cecchini, González, Llamedo, Sánchez, & Rodríguez, 2019; Christensen & Mendoza, 1986; Maassen, 2004):

$$RCI = \frac{D_i}{\sqrt{\left(S_{\text{pre}}\sqrt{1-R_{\text{pre-post}}}\right)^2 + \left(S_{\text{post}}\sqrt{1-R_{\text{pre-post}}}\right)^2}}$$

Ferrer and Pardo (2014) showed that the best false positive rate is obtained when the reliability is estimated with the pre-post correlation ($R_{\text{pre-post}}$).

After applying *SID* and *RCI* to each case in each simulated sample, we declared an individual change to be reliable when its *SID* or *RCI* value was greater than 1.645 (one-sided test) or 1.96 (two-sided test). The post-test mean was expected to be higher than the pre-test mean. Therefore, we adopted a cutoff corresponding to a one-sided right test in the distribution of changes (post-test minus pre-test).

In the single group design, we computed the percentage of reliable improvements for each sample (Percentage of cases with *SID* or *RCI* > 1.645. Two-sided tests used when *ES*=0). In the control-pre-post designs, we calculated in both groups the *percentage of improvements* or P^+ (the percentage of cases with significant "pre<post" differences) and the *percentage of worsenings* or P^- (the percentage of cases with significant "pre>post" differences), and subtracted these to obtain the *net percentage of positive changes*:

$$P_{\text{net}} = (P^+_{\text{exp}} - P^-_{\text{exp}}) - (P^+_{\text{ctrl}} - P^-_{\text{ctrl}}).$$

Finally, with every ES and PIC (200 pairs of values for each simulated condition), we obtained scatterplots to explore the underlying association, and fitted different functions (linear, quadratic, cubic; SPSS *curvefit* procedure) to specify the degree to which the change in ES helps predicting the PIC.

Results

Given the space limitations, only the most relevant results are reported here. Specifically, we report the results based on the *SID* statistic for some representative conditions. The individual-based statistics based on *RCI* yielded very similar results: the correlation between PIC based on *SID* and based on *RCI* was .96. The *mean difference* between them was 0.06, *median* = 0). Results from the rest of conditions and results based on *RCI* are available upon request.

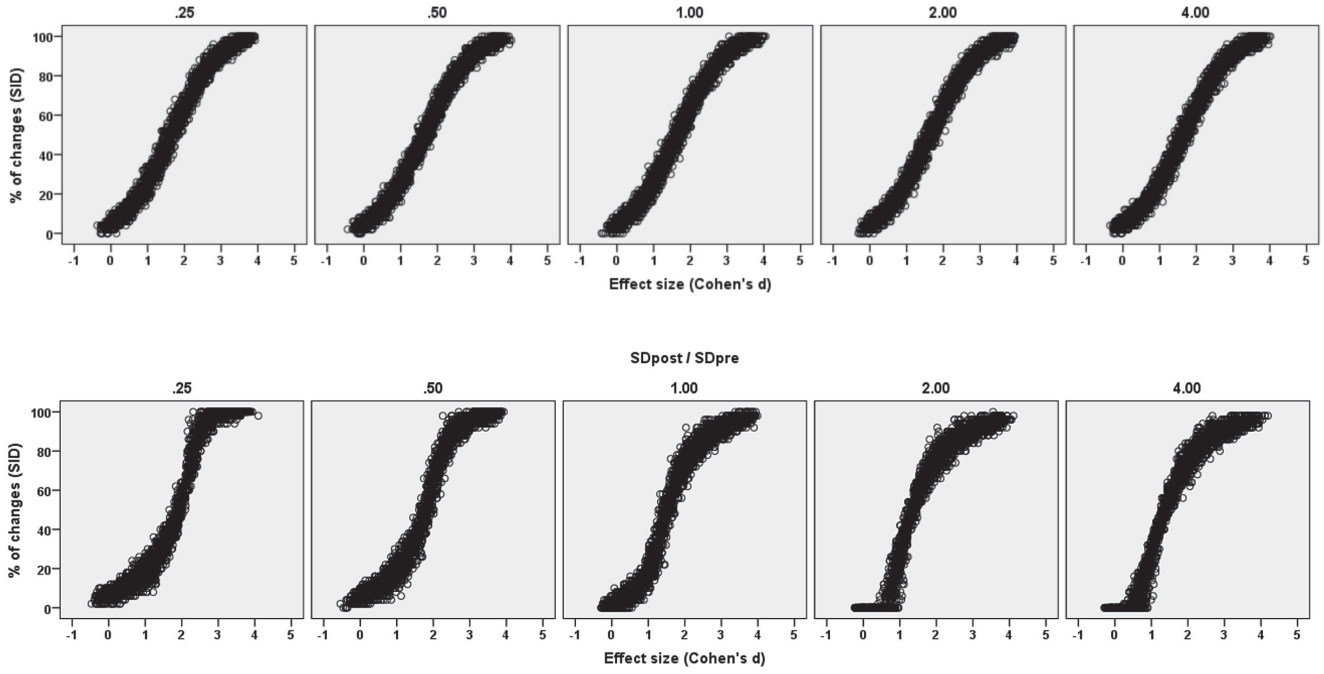

Figure 1. Single group pre-post design. Relation between group effect size (horizontal axis) and percentage of individual changes (vertical axis). Top panel: n=50, R_{XY}=0.50, normal distribution. Bottom panel: n=50, R_{XY}=0.50, extreme negative skewness

Single group pre-post design

We created scatterplots representing the correlation between the ES (Cohen's *d*) and the PIC (computed as the significant *SID* values). Figure 1 shows some of these scatterplots for n=50 and R_{XY}=0.50 (the same trends were observed with n=20 and n=100: greater dispersion with n=20, and smaller dispersion with n=100; different R_{XY} values led to almost identical results). The top panel in Figure 1 depicts one of the scenarios with stronger association (normal distribution); the bottom panel depicts one of the scenarios with weaker association (distributions with greater degree of non-normality). In each scatterplot, each point represents one replication. These replications come from the thirteen effect sizes within a given simulated condition (effects of 13 different sizes and 200 samples per simulated condition: 2,600 points per plot).

To quantify the strength of the underlying association, we fitted linear, quadratic, and cubic functions. In all of them, ES (Cohen's *d* statistic) was used as the independent variable and the PIC (calculated from the *SID* statistic) as the dependent variable. Table 1 reports the coefficient of determination (R^2) from these functions in some representative simulated conditions. These results indicate that all three functions achieved a very good fit. Although fit was slightly better for more complex functions, the R^2 values obtained with the linear function are quite similar to those obtained with the others. Different R_{XY} values led to almost identical results. They are available upon request.

Tables 2, 3 and 4 offer a summary of the results obtained (minimum, medium and maximum R^2 values) segmented by *sample size* (Table 2), degree of *pre-post correlation* (Table 3) and degree of *post-test vs. pre-test variability* (Table 4). First, these results indicate that fit generally improves as the sample size increases; however, this increase is quite small: between n=20 and n=100, the mean R^2 value increases in .04 points for linear and cubic functions, and .03 points for quadratic functions (see Table 2). Second, R^2 does not appear to be altered by the pre-post correlation (see Table 3). Third, and key to the present study, R^2 is ***not*** markedly altered when the post-test variability changes (see Table 4).

Table 1
Single group pre-post design. Fit of the linear, quadratic and cubic functions, for different distribution shapes, sample sizes, and change in standard deviation from pre- to post-test. Predictor: Cohen's d; dependent variable: % of changes based on SID. R_{XY}=0.50

		n=20					n=50					n=100				
		Sd_{post}/Sd_{pre}					Sd_{post}/Sd_{pre}					Sd_{post}/Sd_{pre}				
Shape	**Function**	**0.25**	**0.5**	**1**	**2**	**4**	**0.25**	**0.5**	**1**	**2**	**4**	**0.25**	**0.5**	**1**	**2**	**4**
Skew=-3 Kurt=18	Linear	0.91	0.92	0.91	0.88	0.90	0.92	0.93	0.93	0.90	0.91	0.93	0.94	0.93	0.90	0.92
	Quadratic	0.92	0.92	0.93	0.93	0.94	0.93	0.93	0.94	0.94	0.95	0.93	0.94	0.95	0.94	0.96
	Cubic	0.96	0.96	0.96	0.94	0.95	0.97	0.98	0.97	0.96	0.97	0.98	0.98	0.98	0.97	0.98
Skew=-2 Kurt=9	Linear	0.93	0.94	0.93	0.92	0.92	0.95	0.95	0.95	0.93	0.94	0.95	0.96	0.95	0.94	0.94
	Quadratic	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.97	0.96	0.96	0.97	0.97	0.97
	Cubic	0.97	0.97	0.97	0.96	0.97	0.98	0.99	0.98	0.98	0.98	0.99	0.99	0.99	0.98	0.99
Skew=-1 Kurt=2	Linear	0.95	0.95	0.94	0.94	0.94	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.96	0.96
	Quadratic	0.95	0.95	0.95	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98
	Cubic	0.97	0.97	0.97	0.97	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Skew=0 Kurt=0	Linear	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	Quadratic	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	Cubic	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	0.99	0.99
Skew=1 Kurt=2	Linear	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.97
	Quadratic	0.96	0.96	0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.97	0.97	0.97
	Cubic	0.97	0.97	0.98	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Skew=2 Kurt=9	Linear	0.92	0.93	0.94	0.93	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.96	0.95	0.96
	Quadratic	0.95	0.94	0.94	0.93	0.94	0.96	0.96	0.95	0.95	0.95	0.97	0.96	0.96	0.95	0.96
	Cubic	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Skew=3 Kurt=18	Linear	0.89	0.91	0.92	0.90	0.91	0.91	0.92	0.93	0.92	0.93	0.91	0.93	0.94	0.92	0.93
	Quadratic	0.93	0.93	0.92	0.91	0.92	0.95	0.94	0.93	0.92	0.93	0.95	0.95	0.94	0.92	0.94
	Cubic	0.95	0.96	0.96	0.95	0.96	0.97	0.97	0.98	0.97	0.98	0.97	0.98	0.98	0.97	0.98

Table 2
Single group pre-post design. R^2 values for each function and sample size

	Function	**Min R^2**	**Max R^2**	**Mean R^2**
n=20	Linear	0.88	0.96	0.93
	Quadratic	0.90	0.97	0.95
	Cubic	0.90	0.98	0.95
n=50	Linear	0.91	0.96	0.95
	Quadratic	0.92	0.98	0.96
	Cubic	0.92	0.98	0.96
n=100	Linear	0.94	0.98	0.97
	Quadratic	0.96	0.99	0.98
	Cubic	0.97	0.99	0.99

Table 3
Single group pre-post design. R^2 values for each function and pre-post correlation (R_{XY})

R_{XY}	Function	Min R^2	Max R^2	Mean R^2
0.30	Linear	0.91	0.98	0.94
	Quadratic	0.92	0.98	0.95
	Cubic	0.95	0.99	0.98
0.50	Linear	0.90	0.98	0.94
	Quadratic	0.92	0.98	0.95
	Cubic	0.95	0.99	0.98
0.70	Linear	0.91	0.98	0.95
	Quadratic	0.93	0.98	0.96
	Cubic	0.95	0.99	0.98
0.90	Linear	0.91	0.98	0.94
	Quadratic	0.93	0.98	0.96
	Cubic	0.96	0.99	0.98

Table 4
Single group pre-post design. R^2 values for each function and change in standard deviation from pre- to post-test

Sd_{post}/Sd_{pre}	Function	Min R^2	Max R^2	Mean R^2
0.25	Linear	0.88	0.98	0.94
	Quadratic	0.92	0.98	0.95
	Cubic	0.94	1.00	0.98
0.50	Linear	0.88	0.98	0.95
	Quadratic	0.93	0.98	0.96
	Cubic	0.94	1.00	0.98
1.00	Linear	0.89	0.98	0.95
	Quadratic	0.91	0.98	0.96
	Cubic	0.95	1.00	0.98
2.00	Linear	0.89	0.98	0.94
	Quadratic	0.92	0.98	0.96
	Cubic	0.95	1.00	0.98
4.00	Linear	0.89	0.98	0.94
	Quadratic	0.92	0.98	0.96
	Cubic	0.95	1.00	0.98

With the quadratic function, the R^2 values range between 0.91 and 0.99. With the cubic function, they range between 0.92 and 0.99. Though both functions generally offer a slightly better fit than the linear function, the difference between them makes us think that the linear function is adequate to represent the underlying association: the R^2 values obtained with the linear function were always above 0.89 (found only with n=20 and extreme skewness). In most conditions, they range between 0.90 and 0.98.

Furthermore, the three functions yield very similar predictions. Table 5 contains the regression coefficients obtained for each of them, averaged for conditions. Using these average values, an ES of, for example, d=1, leads to a predicted PIC of 30% (linear and quadratic functions), and 25% (cubic). Though the estimated value for theses coefficients varies slight depending on the simulated condition, the variability is very small: across all conditions, the lowest and highest predictions are 27.2% and 31.9%, respectively. The coefficients in Table 5 allow computing a point estimate for any given ES, aggregated for all conditions. Additionally, we created supplemental tables with the PIC values (based on *SID* and *RCI*) for every condition in our study. They include the average PIC and the empirical quantiles 5 and 95 in our simulated samples. This information can be directly interpreted as the point estimate and 90% confidence interval for a range of ESs in every condition. The supplemental tables are available online at https://github.com/EduardoEstradaRs/Psicothema2020-GroupIndivChangeHeterosk

Focusing on the linear functions, the values for B_0 range between 0 and 3; and the values for B_1 between 28 and 32. These values imply that: (a) in the presence of a null effect, the estimated PIC ranges between 0 and 3%, with an average value of 0%; and (b) for each additional point in ES, the estimated PIC increases by 30 points, with minimum and maximum values of 28% and 32% (note that the predictions below zero or above 100 must be replaced with their respective limits).

Control group pre-post design

Figure 2 shows several scatterplots representing the correlation between ES (quantified using the *omega-squared* statistic) and PIC (calculated as the P_{net} statistic). These plots represent the conditions with n=50 and R_{XY}=0.50 (very similar trends are observed with n=20 and n=100, though with greater dispersion with n=20 and less dispersion with n=100. No noteworthy differences were found with other R_{XY} values). The top and bottom panels in Figure 2 depict, respectively, one condition with a strong association (normal distribution), and another with a weak association (large departure from normality). Every point in Figure 2 represents one of the simulated samples (effects of 13 different sizes and 200 samples per simulated condition: 2,600 points per plot), but now each sample, i.e., each point, represents one experimental and one control group.

To capture underlying association in the scatterplots in Figure 2, we fitted linear, quadratic and cubic functions. The *omega-squared* statistic was the independent variable and the net PIC was the dependent variable. Table 6 shows the coefficient of determination (R^2) obtained with these functions in each simulated condition. These results indicate that the three functions yielded a good fit. Again, the fit is slightly better for more complex functions. Different R_{XY} values led to almost identical results. They are available upon request.

Tables 7, 8 and 9 offer a summary of the results for the control group pre-post design (minimum, maximum and average R^2 values) segmented by *sample size* (Table 7), *pre-post correlation* (Table 8) and *post-test vs. pre-test variability change* (Table 9). First, these results indicate that the fit improves when the sample size increases. This increase is slightly greater than for the single group design: between n=20 and n=100, the average R^2 value increases .08 points for the linear and quadratic functions, and .02 points for the cubic function (see Table 7). Second, R^2 appear to be unaffected by the pre-post correlation (Table 8). Third, and key

Table 5
Single group pre-post design. Regression coefficients (standard errors) for linear, quadratic and cubic functions

	B_0	B_1	B_2	B_3
Linear	0.00 (0.02)	0.30 (0.01)		
Quadratic	-0.04 (0.04)	0.36 (0.09)	-0.02 (0.03)	
Cubic	0.03 (0.04)	0.05 (0.14)	0.21 (0.07)	-0.04 (0.01)

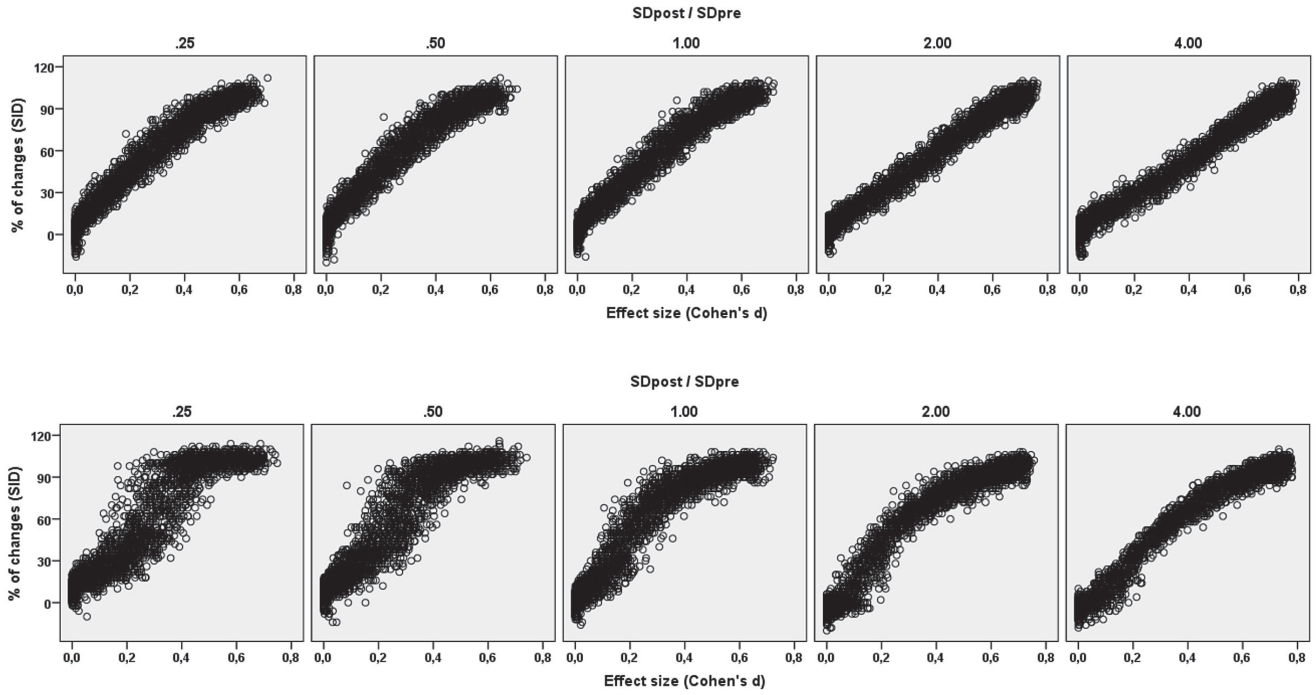

Figure 2. Control group pre-post design. Relation between group effect size (horizontal axis) and percentage of individual changes (vertical axis). Top panel: n=50, R_{XY}=0.50, normal distribution. Bottom panel: n=50, R_{XY}=0.50, extreme negative skewness

Table 6

Control group pre-post design. Fit of the linear, quadratic and cubic functions, for different distribution shapes, sample sizes, and change in standard deviation from pre- to post-test. Predictor: Omega-squared; dependent variable: net % of changes based on *SID*. R_{XY}=0.50

		n=20					n=50					n=100				
		Sd_{post}/Sd_{pre}					Sd_{post}/Sd_{pre}					Sd_{post}/Sd_{pre}				
Shape	**Function**	**0.25**	**0.5**	**1**	**2**	**4**	**0.25**	**0.5**	**1**	**2**	**4**	**0.25**	**0.5**	**1**	**2**	**4**
Skew=-3 Kurt=18	Linear	0.79	0.78	0.81	0.86	0.92	0.87	0.88	0.89	0.92	0.95	0.91	0.92	0.92	0.93	0.96
	Quadratic	0.79	0.81	0.86	0.90	0.92	0.89	0.90	0.95	0.96	0.98	0.92	0.94	0.97	0.98	0.98
	Cubic	0.81	0.82	0.87	0.90	0.93	0.90	0.91	0.95	0.96	0.98	0.94	0.95	0.97	0.98	0.99
Skew=-2 Kurt=9	Linear	0.83	0.83	0.84	0.90	0.93	0.91	0.91	0.92	0.95	0.97	0.95	0.95	0.94	0.96	0.98
	Quadratic	0.83	0.84	0.88	0.92	0.93	0.92	0.93	0.96	0.98	0.98	0.95	0.96	0.98	0.99	0.99
	Cubic	0.85	0.86	0.89	0.93	0.94	0.93	0.94	0.96	0.98	0.98	0.97	0.97	0.98	0.99	0.99
Skew=-1 Kurt=2	Linear	0.86	0.86	0.87	0.91	0.93	0.94	0.93	0.94	0.96	0.97	0.96	0.96	0.96	0.97	0.98
	Quadratic	0.87	0.87	0.90	0.93	0.93	0.95	0.95	0.97	0.98	0.98	0.97	0.97	0.98	0.99	0.99
	Cubic	0.88	0.88	0.90	0.93	0.93	0.95	0.95	0.97	0.98	0.98	0.97	0.98	0.98	0.99	0.99
Skew=0 Kurt=0	Linear	0.88	0.87	0.90	0.93	0.94	0.94	0.94	0.95	0.97	0.97	0.96	0.96	0.97	0.99	0.99
	Quadratic	0.90	0.89	0.91	0.93	0.93	0.96	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.99
	Cubic	0.90	0.90	0.91	0.94	0.94	0.96	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.99	0.99
Skew=1 Kurt=2	Linear	0.84	0.85	0.89	0.93	0.93	0.91	0.92	0.95	0.97	0.96	0.94	0.94	0.97	0.98	0.98
	Quadratic	0.89	0.89	0.89	0.91	0.92	0.96	0.96	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	Cubic	0.89	0.90	0.90	0.93	0.94	0.96	0.96	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.99
Skew=2 Kurt=9	Linear	0.81	0.80	0.85	0.90	0.92	0.87	0.89	0.92	0.95	0.95	0.90	0.92	0.95	0.96	0.96
	Quadratic	0.87	0.86	0.86	0.88	0.91	0.94	0.94	0.93	0.95	0.97	0.97	0.97	0.96	0.97	0.98
	Cubic	0.87	0.86	0.87	0.91	0.94	0.94	0.94	0.94	0.96	0.97	0.97	0.97	0.97	0.98	0.98
Skew=3 Kurt=18	Linear	0.74	0.75	0.81	0.87	0.90	0.82	0.84	0.89	0.91	0.93	0.86	0.89	0.92	0.93	0.93
	Quadratic	0.83	0.82	0.82	0.85	0.89	0.92	0.91	0.90	0.92	0.95	0.95	0.95	0.93	0.94	0.96
	Cubic	0.83	0.82	0.83	0.88	0.93	0.92	0.91	0.91	0.94	0.96	0.95	0.95	0.94	0.96	0.97

for the present study R^2 did not change notably when the post-test variability changed (Table 9). When the standard deviation of the post-test is double or quadruple that of the pre-test, R^2 is altered only slightly: the average R^2 values with the linear function ranged between 0.90 and 0.95.

With the quadratic function, R^2 ranged between 0.78 and 0.99. With the cubic function, R^2 ranged between 0.79 and 0.99. Although both functions generally offered slightly better fit than the linear function, we consider the linear function a better choice to represent the underlying association, given its parsimony and the small difference between the R^2 values: In the least favorable conditions (n=20 and extreme skewness), the lowest R^2 value for the linear function was 0.72. Only 13 of the 420 linear functions (3.1%) yielded R^2 values below 0.80, always with small samples (n=20). In the rest of the simulated conditions, R^2 was never below 0.80. It ranged between 0.80 and 0.90 in 20.5% of the conditions, and above 0.90 in 76.4% of the conditions (reaching R^2=0.99 in some of them).

Table 10 reports the averaged regression coefficients of the functions. These coefficients can be used to estimate the *net* PIC based on the values of the *omega-squared* statistic. The coefficients for the linear function indicate that: (a) with a null effect (*omega-squared*=0), the estimated PIC ranges between 0 and 16.4%, with an average value of 6.64%; and (b) for each 0.10 points of increase in *omega-squared*, the estimated PIC increases by 14.7 points, with minimum and maximum values of 12.7% and 16.5% (the predictions below zero or above 100 must be replaced with their respective limits). For the specific estimates in every condition, see the supplementary tables in https://github.com/EduardoEstradaRs/Psicothema2020-GroupIndivChangeHeterosk

Table 7
Control group pre-post design. R^2 values for each function and sample size

	Function	Min R^2	Max R^2	Mean R^2
n=20	Linear	0.77	0.94	0.87
	Quadratic	0.78	0.94	0.89
	Cubic	0.86	0.98	0.95
n=50	Linear	0.79	0.98	0.93
	Quadratic	0.86	0.98	0.95
	Cubic	0.87	0.98	0.96
n=100	Linear	0.84	0.99	0.95
	Quadratic	0.90	0.99	0.97
	Cubic	0.92	0.99	0.97

Table 8
Control group pre-post design. R^2 values for each function and pre-post correlation (R_{XY})

R_{XY}	Function	Min R^2	Max R^2	Mean R^2
0.30	Linear	0.77	0.99	0.91
	Quadratic	0.78	0.99	0.93
	Cubic	0.79	0.99	0.94
0.50	Linear	0.74	0.99	0.91
	Quadratic	0.79	0.99	0.93
	Cubic	0.81	0.99	0.94
0.70	Linear	0.76	0.99	0.92
	Quadratic	0.80	0.99	0.94
	Cubic	0.81	0.99	0.94
0.90	Linear	0.76	0.99	0.93
	Quadratic	0.78	0.99	0.94
	Cubic	0.79	0.99	0.95

Table 9
Control group pre-post design. R^2 values for each function for different values of post-test standard deviation

Sd_{post}/Sd_{pre}	Function	Min R^2	Max R^2	Mean R^2
0.25	Linear	0.77	0.99	0.90
	Quadratic	0.78	0.99	0.92
	Cubic	0.79	0.99	0.93
0.50	Linear	0.75	0.98	0.90
	Quadratic	0.81	0.99	0.92
	Cubic	0.82	0.99	0.93
1.00	Linear	0.76	0.97	0.90
	Quadratic	0.78	0.98	0.93
	Cubic	0.79	0.98	0.93
2.00	Linear	0.86	0.99	0.94
	Quadratic	0.85	0.99	0.95
	Cubic	0.88	0.99	0.96
4.00	Linear	0.88	0.99	0.95
	Quadratic	0.87	0.99	0.96
	Cubic	0.91	0.99	0.96

Table 10
Control group pre-post design. Regression coefficients (and standard errors) for linear, quadratic and cubic functions

	B_0	B_1	B_2	B_3
Linear	6.64 (5.23)	147.30 (10.38)		
Quadratic	1.76 (4.32)	213.44 (74.26)	-107.73 (105.65)	
Cubic	2.13 (5.26)	203.16 (117.41)	-74.00 (290.14)	-2.59 (197.76)

Discussion

The first objective of the present study was to extend the results in Estrada et al. (2019) to scenarios with a change of variability between pre- and post-test time points. In other words, we sought to verify whether the change in the center of the distribution (average based effect size, ES) is associated with the percentage of individual changes (recovery percentage PIC) when the variability change across time points. The scatterplots of Figures 1 and 2 show that ES increases are monotonically associated with the PIC, regardless of the shape of the underlying association and the pre-post change in variability. Therefore, the results reported by Estrada et al. (2019) are replicated and extended here, under this broad set of new scenarios: even when a compensation or magnification effect exists –and therefore the distribution has lower or greater variability in the post-test time point– the association between average and individual based change statistics is very strong, and increases monotonically.

Our second objective was to find the function that could represent the existing association between ES and PIC. We found that the linear, quadratic and cubic functions all offer excellent fit. However, we hold that the linear function is the best choice: it offers an excellent fit, while being the most parsimonious, and thereby preferable from an applied perspective (Bentler & Mooijaart, 1989; Steele & Douglas, 2006). Again, this result replicates the findings in Estrada et al. (2019), and extends them to scenarios with a change of variance. Note that, because the dependent variables in our regression models are percentages, a function with asymptotical values (e.g., logistic) would be suitable. However, we decided to apply only polynomials of degree 1, 2 and 3 because a) they are simpler to interpret for applied practitioners, and b) the achieved excellent fit in every condition.

Another relevant finding is that *the slope of the regression line is approximately the same in all of the simulated conditions*. In the single group pre-post designs, we found an average value of 0.30 (ranging between 0.28 and 0.32). This means that, for each point of increase in ES (Cohen's *d*), the linear function estimates an increase of 30 points in PIC (i.e., an increase of .10 in *d* is associated with an increase of 3 points in PIC). In the control-pre-post design, the R^2 values indicate that the linear function is also the best choice to represent the underlying association. Based on the average regression coefficients, for each .10 additional points of *omega-squared*, the estimated PIC increases approximately 15 points.

Tables 5 and 10 provide regression coefficients allowing a general point estimation of the expected PIC given an ES value. The specific PIC expected for every condition and ES (based on the average, quantile 5, and quantile 95, obtained in our simulation) are available online at https://github.com/EduardoEstradaRs/Psicothema2020-GroupIndivChangeHeterosk. Researchers interested in computing the expected PIC for any specific set of conditions can consult these tables to make their estimate.

Our third objective was to determine under which conditions the association is found. It was verified in all the simulated conditions. However, the degree of fit of the linear function is not identical in all them: in the most favorable conditions (normality and similar variances in the pre- and post-test distributions), R^2 reached values close to 0.99; in the least favorable conditions (extreme skewness), R^2 dropped to 0.88 in the single group pre-post designs and to 0.74 in the control-pre-post design. However, this happened with n=20. The fit improved considerably with larger samples: with n=100, in the single group pre-post designs, R^2 reached 0.99 in the most favorable conditions and did not fall below 0.94 in the least unfavorable ones. In the control-pre-post designs, R^2 reached 0.99 and did not fall below 0.84.

Implications of our findings

In the clinical setting, it is increasingly common that professionals evaluate the effectiveness of their treatments by computing a recovery percentage (e.g., Ogles et al., 2001). Our results indicate that, when this percentage is unknown, a good estimate can be obtained based on the averaged based effect size, which is usually known and reported in previous studies.

Let us illustrate this idea with an example. Macías et al. (2019) applied an intervention for improving several psycho-social variables in public workers. They compared the pre- and post-test scores of 19 cases receiving the intervention and 19 control cases. According to their Table 2, the treated group experienced a significantly greater improvement in the Mental Health scale, F_{AB} = 45.7, $p < .01$. Based on this information, we can compute $\hat{\omega}^2$ = 1 (45.7 – 1) / (1 (45.7 – 1) + 19) = .37. Now we can use the linear coefficients in our Table 10 to compute the net percentage of changes as *PIC* = 6.64 + 147.3*.37 = 61.19%. This is the net percentage of individual improvements in the treated group.

Our results have a number of methodological implications for the understanding and interpretation of ES measures. For example, meta-analytic studies could include individual based estimations of effect sizes to allow an easier interpretation of the effectiveness of clinical interventions (especially for applied researchers). Future research should examine the sampling distribution and standard errors of the individual based statistics.

In many contexts, it is common to use cutoffs to interpret their magnitude. The best-known and used are those proposed by Cohen (1992) for the typified difference *d* (in single group pre-post designs, 0.20, 0.50 and 0.80 for small, medium and large effect sizes, respectively). These cutoffs were based on the degree of overlap between two normal distributions as the mean difference is changed. Applying these cutoffs to the data simulated in our study leads to 9%, 15% and 24% of individual changes. In the context of a control group pre-post design, similar cutoffs have been proposed for the omega-squared statistic: 0.01, 0.06 and 0.15 for small, medium and large effect sizes, respectively (Kirk, 2013). Applying these cutoffs to the data in our study leads to 8.1%, 15.5% and 27.3% of changes. The estimated PIC is quite similar in both designs: a large ES is associated with approximately 25% of changes. In our view, it is unreasonable to declare that an intervention leading to a reliable change in one out of four patients had a "large effect". Therefore, it appears that the cutoffs habitually used to evaluate effect sizes are, in addition to arbitrary, hardly informative (for further details and examples, see Estrada et al., 2019).

Final remarks

The present study extends the current knowledge about the association between average and individual change statistics. Based on our results, we can generalize the strong association patterns found in previous reports to an even broader set of pre-post scenarios. Specifically, we found that a linear function can be assumed in the presence of changes of variance between the pre- and post- time points. This finding is extremely important, because changes in variance are expected (and often found) in pre-post studies. Particularly, clinical and cognitive interventions –among many others– often lead to individual differences in change that are associated with the baseline level of the participants. Our findings provide evidence for the idea that the average and individual based statistics are strongly associated also in these frequent empirical scenarios.

Acknowledgements

This study was funded by the research project PSI2015-67286-P of the Ministry of Economy and Competitiveness of Spain.

References

Beaton, D. E., Bombardier, C., Katz, J. N., & Wright, J. G. (2001). A taxonomy for responsiveness. *Journal of Clinical Epidemiology, 54*(12), 1204-1217. https://doi.org/10.1016/S0895-4356(01)00407-3

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106*(2), 315-317. https://doi.org/10.1037/0033-2909.106.2.315

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(2), 78-84. https://doi.org/10.1027/1614-2241/a000057

Cecchini, J. A., González, C., Llamedo, R., Sánchez, B., & Rodríguez, C. (2019). The impact of cooperative learning on peer relationships, intrinsic motivation and future intentions to do sport. *Psicothema, 31*(2), 163-169. https://doi.org/10.7334/psicothema2018.305

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17*(3), 305-308. https://doi.org/10.1016/S0005-7894(86)80060-0

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A), 20*(5), 755-762. https://doi.org/10.1076/jcen.20.5.755.1132

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology, 56*(5), 395-407. https://doi.org/10.1016/S0895-4356(03)00044-1

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes, 4*(1), 54. https://doi.org/10.1186/1477-7525-4-54

Estrada, E., Ferrer, E., & Pardo, A. (2019). Statistics for evaluating pre-post change: Relation between change in the distribution center and change in the individual scores. *Frontiers in Psychology, 9*, 2696. https://doi.org/10.3389/fpsyg.2018.02696

Ferrer, R., & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment, 26*(2), 370-383. https://doi.org/10.1037/a0035419

Foster, J. L., Harrison, T. L., Hicks, K. L., Draheim, C., Redick, T. S., & Engle, R. W. (2017). Do the effects of working memory training depend on baseline ability level? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1677-1689. https://doi.org/10.1037/xlm0000426

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed). New York: Routledge.

Hays, W. L. (1988). *Statistics* (2a ed.). Chicago, IL.: Holt, Rinehart and Winston.

Jacobson, N. S., & 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. https://doi.org/10.1037/0022-006X.59.1.12

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed). Thousand Oaks: Sage Publications.

Lemieux, J., Beaton, D. E., Hogg-Johnson, S., Bordeleau, L. J., & Goodwin, P. J. (2007). Three methods for minimally important difference: No relationship was found with the net proportion of patients improving. *Journal of Clinical Epidemiology, 60*(5), 448-455. https://doi.org/10.1016/j.jclinepi.2006.08.006

Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society, 10*(6), 888-893. https://doi.org/10.1017/S1355617704106097

Macías, J., Valero-Aguayo, L., Bond, F. W., & Blanca, M. J. (2019). The efficacy of functional-analytic psychotherapy and acceptance and commitment therapy (FACT) for public employees. *Psicothema, 31*(1), 24-29. https://doi.org/10.7334/psicothema2018.202

Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution and anchor-based approaches in interpretation of changes in Health-Related Quality of Life. *Medical Care, 39*(10), 1039-1047. https://doi.org/10.1097/00005650-200110000-00002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review, 21*(3), 421-446. https://doi.org/10.1016/S0272-7358(99)00058-6

Osma, J., Sánchez-Gómez, A., & Peris-Baquero, Ó. (2018). Applying the unified protocol to a single case of major depression with schizoid and depressive personality traits. *Psicothema, 30*(4), 364-369. https://doi.org/10.7334/psicothema2018.41

Pardo, A., & Ferrer, R. (2013). Clinical significance: False positives in the estimation of individual change. *Anales de Psicología, 29*(2), 301-310. https://doi.org/10.6018/analesps.29.2.139031

Payne, R. W., & Jones, H. G. (1957). Statistics for the investigation of individual cases. *Journal of Clinical Psychology, 13*(2), 115-121. https://doi.org/10.1002/1097-4679(195704)13:2<115::AID-JCLP2270130203>3.0.CO;2-1

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, N.J: Lawrence Erlbaum Associates.

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208-225. https://doi.org/10.1037/met0000126

Schmitt, J. S., & Di Fabio, R. P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *Journal of Clinical Epidemiology, 57*(10), 1008-1018. https://doi.org/10.1016/j.jclinepi.2004.02.007

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smoleń, T., Jastrzebski, J., Estrada, E., & Chuderski, A. (2018). Most evidence for the compensation account of cognitive training is unreliable. *Memory & Cognition, 46*(8), 1315-1330. https://doi.org/10.3758/s13421-018-0839-z

Steele, A. G., & Douglas, R. J. (2006). Simplicity with advanced mathematical tools for metrology and testing. *Measurement, 39*(9), 795-807. https://doi.org/10.1016/j.measurement.2006.04.010

Vindras, P., Desmurget, M., & Baraduc, P. (2012). When one size does not fit all: A simple statistical method to deal with across-individual variations of effects. *PLoS ONE, 7*(6), e39059. https://doi.org/10.1371/journal.pone.0039059