frontiers in Psychology

ORIGINAL RESEARCH
published: 08 January 2019
doi: 10.3389/fpsyg.2018.02696

Statistics for Evaluating Pre-post Change: Relation Between Change in the Distribution Center and Change in the Individual Scores

Eduardo Estrada[1,2]*, *Emilio Ferrer*[2] and *Antonio Pardo*[1]

[1] Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain, [2] Department of Psychology, University of California, Davis, Davis, CA, United States

In a number of scientific fields, researchers need to assess whether a variable has changed between two time points. Average-based change statistics (*ABC*) such as Cohen's *d* or Hays' ω^2 evaluate the change in the distributions' center, whereas Individual-based change statistics (*IBC*) such as the Standardized Individual Difference or the Reliable Change Index evaluate whether each case in the sample experienced a reliable change. Through an extensive simulation study we show that, contrary to what previous studies have speculated, *ABC* and *IBC* statistics are closely related. The relation can be assumed to be linear, and was found regardless of sample size, pre-post correlation, and shape of the scores' distribution, both in single group designs and in experimental designs with a control group. We encourage other researchers to use *IBC* statistics to evaluate their effect sizes because: (a) they allow the identification of cases that changed reliably; (b) they facilitate the interpretation and communication of results; and (c) they provide a straightforward evaluation of the magnitude of empirical effects while avoiding the problems of arbitrary general cutoffs.

Keywords: individual reliable change, effect size estimation, assessment of change, Reliable Change Index (RCI), pre-post change

OPEN ACCESS

Edited by:
Guido Alessandri,
Sapienza University of Rome, Italy

Reviewed by:
Patrizio E. Tressoldi,
Università degli Studi di Padova, Italy
Kane Meissel,
The University of Auckland, New Zealand

****Correspondence:***
Eduardo Estrada
eduardo.estrada.rs@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *28 June 2018*
Accepted: *14 December 2018*
Published: *08 January 2019*

Citation:
Estrada E, Ferrer E and Pardo A (2019) Statistics for Evaluating Pre-post Change: Relation Between Change in the Distribution Center and Change in the Individual Scores. Front. Psychol. 9:2696. doi: 10.3389/fpsyg.2018.02696

The evaluation of change is a key goal in all sciences. In psychological, education, and medical sciences, there is a long tradition of using effect size measures both to quantify the amount of change experienced by a group across several time points, and for comparing such change in multiple groups (Cohen, 1988; Richardson, 1996; Fritz et al., 2012; Grissom and Kim, 2012; Kelley and Preacher, 2012; Pek and Flora, 2018). This paper examines the assessment of change through individuals' responses to standardized tests. Specifically, we focus on situations in which the same variable is measured at two time points in all the individuals in the sample (i.e., *Pre-post research designs*).

Pre-post designs are often used when an intervention is applied between the two time points. Whether the observed change can be attributed to the intervention or not depends on a number of factors, including whether (a) a control group exists; (b) the study is experimental, quasi-experimental or observational; (c) relevant covariates and cofounds have been adequately controlled (Fisher, 1935; Rubin, 1974; Shadish et al., 2002; Pearl, 2009; Mayer et al., 2016).

Very often, researchers want to cause a change with their interventions. Some examples are: (a) A school teacher applying a visual-spatial training program to children of ages 10–12 would want them to increase their ability (c.f. Lowrie et al., 2017); (b) Numerous programs for cognitive training are intended to increase working memory capacity—and, ultimately, general cognitive ability—in their participants (c.f. Jaeggi et al., 2008); (c) Interventions with teenagers with autism spectrum disorders typically aim at improving their interpersonal and communication skills, among others; (d) Interventions in clinical psychology typically are intended to change the clients' behavior so they can adapt better to their environment and increase their quality of life (e.g., gain social skills, control their anger or anxiety, improve their depressive symptoms, or avoid their maladaptive behaviors, among others; c.f. Muroff et al., 2014); (e) A pharmacological treatment for obesity will be successful if the patients reduce their weight (c.f. Pi-Sunyer et al., 2015)[1].

In a pre-post research design, some criterion is needed to determine large or small change. Here we focus on *distribution based methods* (i.e., there is no external information or clinical referents, other than the test scores; Lydick and Epstein, 1993; Crosby et al., 2003; Revicki et al., 2008). These methods attempt to identify the smallest change that cannot be explained by sampling random fluctuations or by measurement error (Jacobson and Truax, 1991; Crosby et al., 2003; Bauer et al., 2004). This amount of change is usually called *statistically reliable*, *minimally detectable* or just *reliable change* (Maassen, 2000; Beaton et al., 2001; de Vet et al., 2006).

To detect a reliable change, two approaches can be adopted. We termed them the *average-based change* approach (ABC) and the *individual-based change* approach (IBC). The aim of ABC is to evaluate whether a group, as a whole, experienced a reliable change. In turn, the goal of IBC is to identify specific individuals who showed change. To assess ABC, researchers often use a statistic that describes the center of the distributions (often, the pre and post means), by using null hypothesis tests and effect size measures (c.f., Cohen, 1988; Fritz et al., 2012; Grissom and Kim, 2012; Pek and Flora, 2018). To assess IBC, researchers may use various indices that can be grouped under the name of *reliable change indices*. Some of these indices are based on standardization of pre-post differences, others on the standard error of measurement, and yet others on linear regression predictions (Crosby et al., 2003; Ferrer and Pardo, 2014).

The goal of this paper is twofold. First, we want to investigate the relation between ABC and IBC statistics, and to describe such a relation in mathematical terms. We show that, contrary to what other previous studies have speculated, both approaches are strongly related. Second, we attempt to draw researchers' attention to a set of tools derived from individual-based statistics. These are simple tools that can provide help in a variety of research contexts. We show how they can be used for intuitive interpretation and communication of research results, and how they can replace arbitrary cutoffs (e.g., Cohen, 1988) commonly used for deciding when an effect is "small" or "large."

ARE THE AVERAGE-BASED AND THE INDIVIDUAL-BASED APPROACHES RELATED?

Many studies have argued that the information provided by these two approaches is different. Below are some examples:

> "Statistical methods based on the General(ized) Linear Model (…) have optimal power when individuals behave identically (…). When there exists genuine, idiosyncratic variations in the effect of a factor, (…) the effect of a factor can be significant for every individual (…) while Student and Fisher tests yield a probability close to one if the population average is small enough" (Vindras et al., 2012, p. 2).

> "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" (Schmitt and Di Fabio, 2004, pp. 1008–1009).

Similar ideas can be found in other studies (e.g., Ottenbacher et al., 1988; Testa, 2000). Accordingly, it appears that average and individual approaches focus on *different* aspects of change, inasmuch as knowing that the center of the scores distribution changed provides no information about which particular individuals changed. Indeed, the change in the distribution center and the percentage of individual changes are calculated in very different ways.[2]

However, it is not evident whether these two approaches are completely *independent*. Rather, it is reasonable to think that the larger the displacement of distribution center, the higher the percentage of reliable individual changes. In fact, the higher the mean of the pre-post differences, the more likely it is that a pre-post difference exceeds a certain cutoff. For example, if the pre-post differences distribution is normal, the probability associated with each cutoff is known. If the mean of the differences equals zero, the probability of finding cases above 1.645 standard deviations equals 0.05. If the mean of the differences is 0.5 standard deviations above zero, the probability of finding cases above 1.645 standard deviations equals 0.13, etc. However, these probabilities are unknown when the pre-post differences distribution is *non-normal*, which is the usual case in applied contexts.

One study showed that the pre-post effect size observed (i.e., the magnitude of change in distribution center) is the main determinant of the percentage of individuals showing pre-post change (Norman et al., 2001). This simulation study revealed that the relation between effect size and percentage of change is approximately linear for effect sizes below one, with normal and moderately skewed distributions, and regardless of the cutoff to detect a change. Therefore, at least under certain conditions, the mean change can yield some information

[1] In some cases, interventions aim at decelerating or stopping changes that are already happening: i.e., an intervention for the elderly aiming at stopping or reducing the speed of decline of some cognitive function. In such situations, "no change" is evidence of treatment success.

[2] In computing the *distribution center* all cases are used, each one of them contributing its proportional share of change; in computing the *percentage of changes* only cases above a given cutoff are involved and, moreover, all of them equally weighted regardless of their change.

about the percentage of individual changes. A later study using empirical data found consistent results (Lemieux et al., 2007). However, these papers did not report any mathematical function to estimate the percentage of changes based on the change in the distribution center, nor did they report the fit that such a function may achieve, which would be useful to assess the quality of its estimations.

The scarcity of studies on this topic and the lack of sound conclusions suggest that more research is needed to understand the relation between the change in the distribution center and the percentage of individual changes.

THE PRESENT STUDY

Our first goal in this article was to investigate the relation between ABC and IBC statistics, and to mathematically describe that relation. Specifically, we sought to: (a) investigate whether ABC and IBC are related; (b) if so, identify its shape, a mathematical function that best represents it, and the goodness of fit of such function; and (c) determine what conditions affect the nature of the relation. For this, we conducted a simulation study corresponding to two of the most common designs in the behavioral and social sciences: a "pre-post design" and a "control group pre-post design." To our knowledge, this is the first study applying individual change indices to a pre-post design with a control group. Importantly, we studied this relation in scenarios with both normal and non-normal distributions.

Our second goal was to promote the use of individual-based statistics as a simple and useful tool for addressing important research questions. Based on our simulation results, we show that such statistics can be used to interpret research results and make decisions in applied settings.

METHODS

We simulated data for two scenarios in which the same variable is measured at two time points (e.g., before and after the intervention) for each individual within a group of participants. We generated two different pre-post research designs: *with* and *without a control group*.

Including a single group design is important because: (a) this is a common scenario in applied contexts; and (b) all indices describing the percentage of individual changes were developed for settings with a single treated group (Payne and Jones, 1957; Jacobson and Truax, 1991; Crawford et al., 1998; Hageman and Arrindell, 1999; Wyrwich et al., 1999). On the other hand, it is well known that including a control group—ideally, with random assignment—provides stronger evidence for attributing the change to the treatment (Shadish et al., 2002; Feingold, 2009).

Simulation Conditions

To define the simulation conditions, we manipulated four criteria (for a summary, see **Table 1**):

a. *Effect size* in the experimental/treatment group ($\delta_{\text{exp}} = \mu_{\text{dif.exp}}/\sigma_{\text{dif.exp}}$). We computed the effect size as the standardized mean of the pre-post differences (Cohen, 1988; see the discussion and Appendix 1 in **Supplementary Data Sheet 2** for considerations about using a different standardizer). We chose 13 effect sizes ranging from 0 to 3.6 with 0.3 point increases (e.g., an effect size of 0.6 indicates that the mean of the pre-post differences $\mu_{\text{dif.exp}}$ is 0.6 times the standard deviation of the individual pre-post differences $\sigma_{\text{dif.exp}}$). The rationale for choosing this wide range of effects, from a null effect to an extremely large one, was to allow the percentage of individual changes to comprise its full range (0–100%). In our analyses we assumed that the mean scores increased over time. To calculate the differences, we subtracted pre-test score from the post-test score. Consequently, because we generated positive effects in our simulation, we used right one-tailed tests.

   In the single group pre-post design, we generated data for the treatment group only. In the control group pre-post design, we added data for a control group with no expected pre-post mean differences, (i.e., $\delta_{\text{ctrl}} = 0$). The values for the rest of simulation criteria were the same for the control and treatment groups in every conditions (see below).

   Importantly, this value was the *mean* effect size in the population. Centered on this mean, a random distribution of individual changes was created, and each case within the sample experienced a different amount of change. The variance of this distribution depended on the pre-post correlation (see point *c* below). **Figure 1** depicts the pre, post and change scores for one sample.

b. *Sample size of each group* (*n*). We chose three sample sizes (25, 50, and 100) to simulate what is usually considered small, medium, and large sample sizes in clinical work (Crawford and Howell, 1998). In the control group design, both groups had the same sample size.

c. *Pre-post correlation* ($\rho_{\text{pre}-\text{post}}$): 0.5, 0.7, and 0.9. We chose these values to simulate a range of common correlations in applied settings (Pedhazur and Schmelkin, 1991; Nunnally and Bernstein, 1994. Note that correlations <0.5 are very uncommon in repeated measures settings). We used the Pearson's correlation coefficient. In the control group design, both groups were expected to have the same correlation value. With $\sigma_{\text{pre}} = \sigma_{\text{post}} = 1$, these three values lead to a standard deviation of the differences (σ_{dif}) of 1, 0.775, and 0.447 respectively—i.e., higher pre-post correlation entails lower variance of the differences (See Appendix 3 in **Supplementary Data Sheet 2** for a discussion on the effect of measurement error).

d. *Shape of the pre and post distributions*. Given that moderate and severe deviations from normality are often found in applied contexts (Micceri, 1989; Blanca et al., 2013), we simulated seven different conditions by modifying the degree of skewness (g_1) and kurtosis (g_2): (1) extreme negative skewness: $g_1 = -3, g_2 = 18$; (2) moderate negative skewness: $g_1 = -2, g_2 = 9$; (3) mild negative skewness: $g_1 = -1, g_2 = 2$; (4) normality: $g_1 = 0$, $g_2 = 0$; (5) mild 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 kurtosis is partially conditioned by the skewness. Less than 5% of real data is expected to have more extreme distributions (Blanca et al., 2013). In the control group design, both groups

TABLE 1 | Summary of simulation conditions and computed statistics.

<table>
<tr><th></th><th colspan="4">Single group design</th><th colspan="4">Control group design</th></tr>
<tr><td colspan="9">SIMULATION CONDITIONS</td></tr>
<tr><td>Effect size in the experimental group</td><td colspan="4">$\delta_{exp} = \mu_{dif.exp}/\sigma_{dif.exp} = \{0 \text{ to } 3.6\}$ in 0.3 steps</td><td colspan="4">$\delta_{exp} = \mu_{dif.exp} / \sigma_{dif.exp} = \{0 \text{ to } 3.6\}$ in steps of 0.3
$\delta_{ctr} = \mu_{dif.ctr} / \sigma_{dif.ctr} = 0$</td></tr>
<tr><td>Sample size</td><td colspan="4">$n_{exp} = \{25, 50, 100\}$</td><td colspan="4">$n_{exp} = n_{ctr} = \{25, 50, 100\}$</td></tr>
<tr><td>Pre-post correlation</td><td colspan="4">$\rho_{pre-post,exp} = \{0.5,0.7,0.9\}$</td><td colspan="4">$\rho_{pre-post,exp} = \rho_{pre-post,ctr} = \{0.5,0.7,0.9\}$</td></tr>
<tr><td>Shape of the pre and post distributions (equal for pre and post and for both groups)</td><td>Skew:
Kurt:</td><td>−3
18</td><td>−2
9</td><td>−1
2</td><td>0
0</td><td>1
2</td><td>2
9</td><td>3
18</td></tr>
<tr><td>Average-based change statistic</td><td colspan="4">$d = (M_{post}\text{-}M_{pre})/S_{dif}$</td><td colspan="4">$\widehat{\omega}^2 = \frac{gl_{AB}(F_{AB}-1)}{gl_{AB}(F_{AB}-1)+N}$</td></tr>
<tr><td>Individual-based statistic (based on <i>SID</i> or <i>RCI</i>)</td><td colspan="4">Percentage of reliable improvements</td><td colspan="4">$P_{net} = \left(P^{+}_{\mathrm{exp}} - P^{-}_{\mathrm{exp}}\right) - \left(P^{+}_{\mathrm{ctr}} - P^{-}_{\mathrm{ctr}}\right)$</td></tr>
</table>

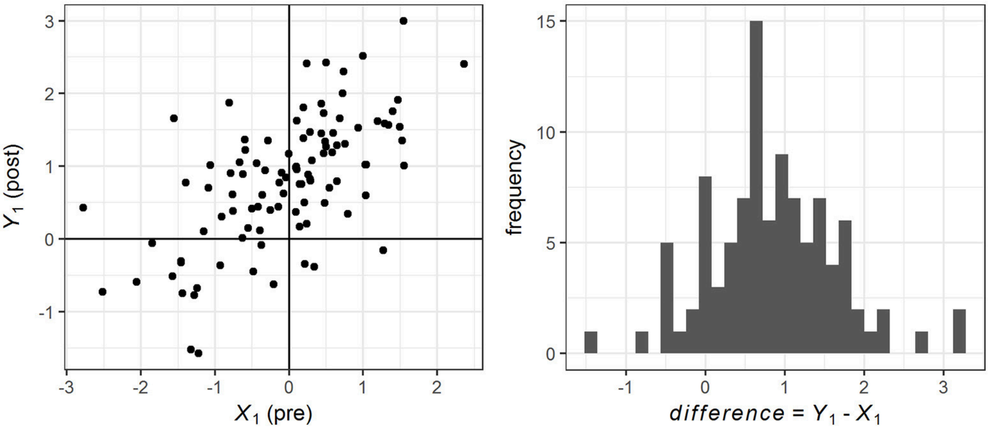

FIGURE 1 | Pre, post and difference scores for one sample of $n = 100$, with $\delta_{exp} = 1.2$, $\rho_{pre-post} = 0.7$, and normal distribution. Note that the amount of change is different for every individual.

were expected to have the same shape for the pre- and post-distributions.

Simulation Procedure

By combining the four criteria described above we generated 13 × 3 × 3 × 7 = 819 different conditions for the simulation. For each of these conditions, we generated 500 samples (409,500 samples in total). This was done separately for the simple pre-post design (one experimental group per sample) and for the control group pre-post design (one experimental and one control group per sample). We used MatLab 2011a to perform the simulation. The code is available in the **Supplementary Data Sheet 1**.

In the single group design, we first generated a matrix $\mathbf{X_1} = (X_1{}^*, Y_1{}^*)$ containing n pairs of scores in two non-correlated variables. Scores were generated by using Pearson's distribution system. Both variables had the same mean, standard deviation, skewness, and kurtosis. The mean was always fixed to zero and the standard deviation was fixed to one. Skewness and kurtosis were systematically modified according to g_1 and g_2 values explained previously. X and Y were generated randomly to ensure that the post score may be the same, higher, much higher, lower or much lower than their corresponding pre-score, as is typically the case in real data.

Second, we fixed the correlation value between variables in $\mathbf{X_1}$ by applying the Cholesky covariance decomposition of correlation matrix $\mathbf{R}$ corresponding to the chosen correlation value ($\rho_{pre-post}$). The resulting matrix $\mathbf{M_1} = (X_1, Y_1)$ contained two variables (X_1 = pre; Y_1 = post) with skewness, kurtosis and ρ_{XY} (or $\rho_{pre-post}$) values similar to the specified ones. This transformation ensured that the post-scores were not independent of the pre-scores, as is also the case in real data. Note that, although simulating the difference scores would be simpler and faster than simulating pre and post scores, it would make it impossible to study the effect of the pre-post correlation.

In the last step we modified Y_1 to adapt it to the desired mean value in each condition. For this purpose, we added the standard deviation of pre-post differences, multiplied by the corresponding value of δ_{exp}, to each individual Y_1 value.

In the control group design, the procedure was identical except for two changes: (a) instead of only one matrix in each replication, we generated a pair of independent matrices

$\mathbf{X_1} = (X_1{}^*, Y_1{}^*)$ and $\mathbf{X_2} = (X_2{}^*, Y_2{}^*)$ for simulating the scores of the experimental ($\mathbf{X_1}$) and control ($\mathbf{X_2}$) groups; and (b) we modified Y_1 in the experimental matrix only to adapt it to the desired mean value in each condition (whereas the mean for Y_2 was not changed for the control group).

Importantly, this procedure ensured that every case experienced a different amount of change. **Figure 1** depicts pre-, post- and difference scores for one sample of $n = 100$, with $\delta_{\text{exp}} = 1.2$, $\rho_{\text{pre}-\text{post}} = 0.7$, and normal distribution.

Data Analysis

In the single group pre-post design, we computed the empirical *group* or *average change* for each sample by calculating the difference between the post- and the pre-test means, and dividing such difference by the standard deviation of the differences,

$$d = (M_{\text{post}} - M_{\text{pre}})/S_{\text{dif}} \tag{1}$$

In this paper we use d to refer to the result of applying Equation 1. See the discussion and Appendix 1 in **Supplementary Data Sheet 2** for a discussion on a different computation of the standardized mean difference.

In the control group pre-post design, we quantified the average change by using the ω^2 statistic associated with the interaction between the between-subjects factor A (group) and the within-subjects factor B (pre- and post-test). The net change is captured by comparing the pre-post change in the experimental group with the pre-post change in the control group (Hays, 1988; Kirk, 2013). For our design, ω^2 can be estimated as

$$\omega^2 = \frac{gl_{AB}\,(F_{AB} - 1)}{gl_{AB}\,(F_{AB} - 1) + N},$$

were F_{AB} is the interaction F statistic, gl_{AB} are the interaction degrees of freedom, and N is the total number of scores in the design (adding both groups).

To identify which individual scores showed a reliable change (i.e., which cases fell above a certain cutoff after being standardized) and then calculate the percentage of individual changes for each sample, we decided to use two *individual change* indices. We chose two indices that have shown lowest false negative rates (see Ferrer and Pardo, 2014).

a. *Standardized individual difference* (*SID;* Payne and Jones, 1957). The standardized score resulting from dividing the individual pre-post difference (D_i) by the standard deviation of these differences (S_{dif}), as

$$SID = D_i/S_{\text{dif}}.$$

This standardization was proposed to assess the degree of discrepancy between two scores (Payne and Jones, 1957). If the distribution of pre-post difference is normal, 95% of *SID* will fall between ± 1.96 values, and 90% between ± 1.645 values.

b. *Reliable Change Index* (*RCI*; Jacobson et al., 1984, 1999; Jacobson and Truax, 1991). This is probably the most popular individual change index. It is based on the standard error of measurement. Of the several available versions, we used one in which the equality of pre- and post-test variances is not assumed (see Christensen and Mendoza, 1986; Jacobson and Truax, 1991; Maassen, 2004). This version is specified as:

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

Using this index, the lower false positive rate is achieved when reliability is estimated from the pre-post correlation ($R_{\text{pre}-\text{post}}$) (Ferrer and Pardo, 2014).

These two indices were computed for each individual case in all the simulated samples. We considered an individual change to be reliable when its corresponding *SID* or *RCI* was higher than 1.96 (two-tailed test) or 1.645 (one-tailed test) points. In the single group pre-post design, we applied one cutoff of 1.645. In the control group pre-post design we performed two-tailed tests (cutoffs of −1.96 and 1.96) for all conditions because the procedure is intended to compare the effectiveness of two different treatments in real scenarios. Hence, it is important to take into consideration the proportion of worsened cases, not only the improved ones.

In the single group design, we computed the percentage of reliable improvements for each sample. In the control group design, we computed the proportion of both worsened (P^-) and improved (P^+) cases in each group within the samples, and then subtracted the result for the control group (*ctrl*) from this same result in the experimental group (*exp*). This procedure yielded a net percentage of positive changes attributable to treatment[3](P_{net}):

$$P_{\text{net}} = \left(P^+_{\text{exp}} - P^-_{\text{exp}}\right) - \left(P^+_{\text{ctr}} - P^-_{\text{ctr}}\right) \tag{2}$$

Then we examined the relation between the change estimated with ABC statistics and the change estimated using IBC statistics by fitting several regression functions.

Finally, with each empirical effect size and percentage of individual changes (500 pairs of values for each condition in the simulation, i.e., a pair by sample), we obtained: (a) a scatterplot to inspect the underlying relation between the two statistics, and (b) several different regression functions to quantify the extent to which the change in the distribution center is predictive of the percentage of individual changes. This was done separately for each research design.

RESULTS

For brevity of presentation, we report here the most representative results. For all conditions in both designs, the properties of the generated samples corresponded to those

[3] Norman et al. (2001), inspired on Guyatt et al. (1998), proposed an alternative *corrected* method for computing P_{net}:

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

Since the results obtained with this equation and with [2] and are nearly identical, here we will only inform about the results obtained with [2].

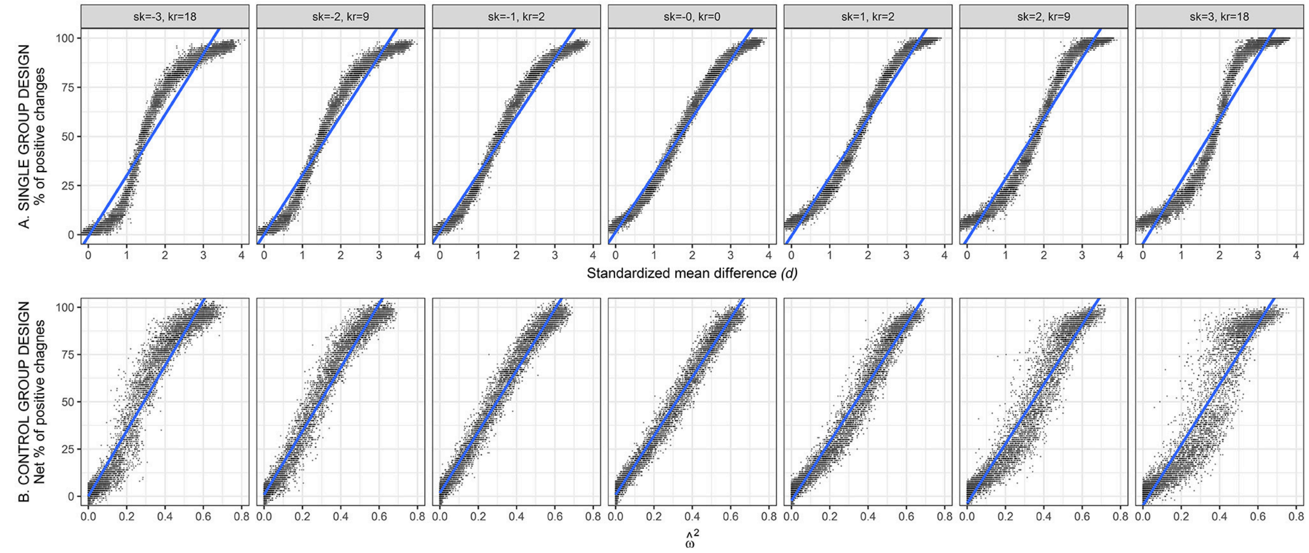

FIGURE 2 | Relation between average-based change (horizontal axis) and individual-based change (vertical axis). Top row **(A)** shows the data for the simple group pre-post design. Bottom row **(B)** shows the data for the pre-post design with a control group. Data based on *SID* with $n = 100$ and $\rho_{\text{pre-post}} = 0.7$. Sk, skewness; Kr, Kurtosis.

imposed in the simulation. We report the results regarding *SID* only; those based on the *RCI* are similar. Results from all conditions and based on the *RCI* are available upon request.

SINGLE GROUP PRE-POST DESIGN

To examine the relation between ABC and IBC, we first plotted the effect size measured by *d* (average-based change) against the percentage of individual changes (individual-based change). **Figure 2** (top row) shows scatterplots based on *SID* index, for $n = 100$ and $\rho_{\text{pre-post}} = 0.7$. Each of the points in these scatterplot depicts one of the simulated samples (i.e., 13 effect sizes × 500 simulated samples = 6,500 points per scatterplot). The patterns with $\rho_{\text{pre-post}} = 0.5$ and $\rho_{\text{pre-post}} = 0.9$ were similar. We report here the conditions with the largest sample size to illustrate the shape of the relation with greater clarity. The same pattern is observed for $n = 25$ and $n = 50$, but with higher variability. In other words, any particular *d* value corresponds to the same mean percentage of changes, but a smaller sample size leads to more scattered points due to higher sampling error.

To quantify the relations detected in **Figure 2**, we estimated four different regression functions: linear, quadratic, cubic, and logistic. In every case, *d* (average-based effect size) was used as the independent variable and the percentage of changes (individual-based effect size) as the dependent variable.

Table 2 reports the coefficient of determination (R^2) for the four functions, for $n = 25$. Because the dispersion in the various scatterplots decreases as sample size increases, these R^2 values were the lowest of all values. Nevertheless, even with $n = 25$, three of the four functions provided an excellent fit. First, the linear function achieves R^2 values around 0.90 in negatively skewed distributions and above 0.90 values in the remaining distributions, reaching 0.96. With $n = 50$ and $n = 100$, R^2 ranges between 0.91 and 0.98; the lowest values are observed in the conditions with more extreme skewness. Second, the quadratic function achieves R^2 values similar to the linear function, although slightly higher in negative skewness conditions. Third, the cubic function yields R^2 values between 0.96 and 0.98, although at the cost of introducing more complexity. Fourth, the logistic function yields the lowest values, between 0.68 and 0.78.

Three of the four adjusted functions offered a very good fit to the data. Moreover, they offered very similar predictions. For example, with $n = 25$, $\rho_{\text{pre-post}} = 0.70$, and $\delta_{\text{exp}} = 1$, the predicted value (the estimated percentage of changes) is 30.7% for the linear function, 31.7% for the quadratic function, and 26.9% for the cubic function. Of these, the linear function is the most parsimonious, especially for applied settings (Bentler and Mooijaart, 1989; Maxwell and Delaney, 2004; Steele and Douglas, 2006). **Table 3** reports the coefficients from the linear function. These coefficients can be used to estimate the percentage of individual changes from the effect size *d*. Given that the value of the former can range from −100 to 100, the constant coefficient B_0 is fairly close to zero in every case (with absolute values ranging from 0.09 to 2.50, and standard errors <0.27; $p > 0.05$ in all cases), and the slope coefficient B_1 is close to 30 (28.75 to 30.86, with standard error < 0.12). Results with other conditions were similar in all regards.

Results from the linear function indicate that: (a) when effect size is zero, the expected percentage of changes (computed using *SID*) ranges between 0 and 3%, and (b) for each extra point of effect size, the expected percentage of changes rises by 30 points. Because prediction is done using percentages, values below zero and above 100 must be replaced by their respective limits.

PRE-POST DESIGN WITH CONTROL GROUP

Figure 2 (bottom row) shows the relation between $\hat{\omega}^2$ (average-based effect size measure) and the net percentage of changes (individual-based effect size measure). The latter was calculated from *SID* ($n = 100$ and $\rho_{\text{pre}-\text{post}} = 0.7$). Each of the points in these scatterplot depicts one of the simulated samples *comprising one control and one experimental group*. As in the top row, we report the results for the conditions with the largest sample size. The smaller sample sizes yielded the same pattern yet with higher variability. Patterns with the other $\rho_{\text{pre}-\text{post}}$ values were similar.

TABLE 2 | R^2 of linear, quadratic, cubic and logistic functions for the single group design.

Distribution	$\rho_{\text{pre-post}}$	Linear	Quadratic	Cubic	Logistic
Sk = −3, Kr = 18	0.5	0.919	0.938	0.965	0.695
	0.7	0.909	0.937	0.961	0.676
	0.9	0.897	0.940	0.956	0.681
Sk = −2, Kr = 9	0.5	0.937	0.951	0.974	0.701
	0.7	0.934	0.954	0.972	0.683
	0.9	0.928	0.955	0.971	0.679
Sk = −1, Kr = 2	0.5	0.951	0.960	0.979	0.705
	0.7	0.951	0.962	0.979	0.687
	0.9	0.946	0.962	0.977	0.679
Sk = 0, Kr = 0	0.5	0.962	0.965	0.982	0.723
	0.7	0.963	0.965	0.982	0.724
	0.9	0.962	0.964	0.982	0.722
Sk = 1, Kr = 2	0.5	0.956	0.956	0.979	0.740
	0.7	0.957	0.957	0.981	0.751
	0.9	0.955	0.956	0.980	0.756
Sk = 2, Kr = 9	0.5	0.943	0.943	0.975	0.752
	0.7	0.944	0.944	0.976	0.756
	0.9	0.939	0.942	0.973	0.770
Sk = 3, Kr = 18	0.5	0.927	0.927	0.969	0.757
	0.7	0.923	0.925	0.967	0.764
	0.9	0.915	0.921	0.960	0.779
Mean value		0.939	0.949	0.973	0.723
Min. value		0.897	0.921	0.956	0.676
Max. value		0.963	0.965	0.982	0.779

n = 25; Independent variable: d; Dependent variable: percentage of changes based on SID; Sk, skewness; Kr, kurtosis.

To quantify the relation observed in the bottom row of **Figure 2**, we estimated four different regression functions: linear, quadratic, cubic and logistic. In every case, $\hat{\omega}^2$ was used separately as the independent variable, and the net percentage of individual changes served as the dependent variable. The four functions were estimated for each of the conditions simulated. **Table 4** reports the coefficient of determination (R^2) for these four functions. These results are based on net percentage of individual changes calculated with *SID* index and $n = 25$. Because dispersion from the various scatterplots decreases as sample size increases, R^2 values from **Table 4** were lower than those achieved with $n = 50$ and $n = 100$.

Overall, the four functions achieved a very good fit. The R^2 values were higher when the distributions approached normality. The quadratic and cubic functions achieved a slightly better fit than the linear function, but only with negative skewness; the logistic and linear functions achieved similar fit. As in the single group design, the linear function was deemed preferable because it is the most parsimonious, with only minimal loss of fit.

In **Table 5** (analogous to **Table 3** in the single group design) we report the coefficients from the linear function with $n = 25$. These coefficients allow estimating the net percentage of individual changes from the effect size measures. The intercept (B_0) ranges from −0.04 to approximately 6, with a mean of 2.41 and standard errors ranging between 0.19 and 0.33. The slope (B_1) ranges from 140 to 165, with a mean of 153 and standard errors ranging between 0.53 and 0.91. As an example, if we consider the results for the normal distributions, these coefficients indicate that for a null effect size ($\hat{\omega}^2 = 0$), the linear function yields an estimated net percentage of changes of approximately 2.5%. For each additional 0.10 points of $\hat{\omega}^2$, the net percentage of changes increases in approximately 15 points (as we are predicting percentages, values beyond zero, and 100 must be replaced by their respective limits). Note that the changes in pre-post correlation do not substantially alter the coefficients B_0 and B_1 in **Table 5**. Similar results were found with the other sample sizes.

TABLE 3 | Coefficients (and standard errors) for the lineal regression model in the single group design.

	$\rho_{\text{pre-post}} = 0.5$		$\rho_{\text{pre-post}} = 0.7$		$\rho_{\text{pre-post}} = 0.9$	
	B_0	B_1	B_0	B_1	B_0	B_1
Sk = −3, Kr = 18	−0.01 (0.24)	30.08 (0.11)	0.48 (0.25)	29.89 (0.12)	1.30 (0.27)	29.69 (0.12)
Sk = −2, Kr = 9	1.11 (0.20)	29.43 (0.09)	1.09 (0.21)	29.51 (0.10)	1.57 (0.22)	29.33 (0.10)
Sk = −1, Kr = 2	1.65 (0.18)	29.07 (0.08)	1.76 (0.18)	29.05 (0.08)	2.15 (0.18)	28.92 (0.09)
Sk = 0, Kr = 0	1.88 (0.15)	28.82 (0.07)	1.89 (0.15)	28.83 (0.07)	1.84 (0.15)	28.89 (0.07)
Sk = 1, Kr = 2	0.42 (0.17)	29.43 (0.08)	0.44 (0.16)	29.46 (0.08)	0.30 (0.17)	29.49 (0.08)
Sk = 2, Kr = 9	−0.49 (0.19)	29.93 (0.09)	−1.02 (0.19)	30.11 (0.09)	−1.12 (0.20)	30.12 (0.09)
Sk = 3, Kr = 18	−1.69 (0.22)	30.52 (0.11)	−2.45 (0.23)	30.78 (0.11)	−2.34 (0.24)	30.68 (0.12)

n = 25; independent variable: d, dependent variable: percentage of changes based on SID. Sk, skewness; Kr, kurtosis.

TABLE 4 | R^2 of linear, quadratic, cubic and logistic functions for the $n = 25$ conditions of the control group pre-post design.

Distribution	$\rho_{\text{pre-post}}$	Linear	Quadratic	Cubic	Logistic
Sk = −3 Kr = 18	0.5	0.857	0.881	0.885	0.858
	0.7	0.857	0.887	0.890	0.858
	0.9	0.841	0.886	0.888	0.843
Sk = −2Kr = 9	0.5	0.894	0.907	0.911	0.894
	0.7	0.891	0.912	0.915	0.892
	0.9	0.887	0.917	0.920	0.888
Sk = −1Kr = 2	0.5	0.916	0.924	0.926	0.916
	0.7	0.916	0.927	0.929	0.916
	0.9	0.916	0.931	0.933	0.917
Sk = 0Kr = 0	0.5	0.926	0.929	0.930	0.926
	0.7	0.928	0.930	0.931	0.927
	0.9	0.924	0.926	0.928	0.924
Sk = 1Kr = 2	0.5	0.906	0.906	0.908	0.905
	0.7	0.902	0.902	0.903	0.901
	0.9	0.898	0.898	0.900	0.898
Sk = 2Kr = 9	0.5	0.866	0.867	0.868	0.867
	0.7	0.865	0.865	0.867	0.865
	0.9	0.841	0.841	0.843	0.842
Sk = 3Kr = 18	0.5	0.824	0.825	0.827	0.825
	0.7	0.815	0.816	0.817	0.816
	0.9	0.785	0.786	0.787	0.785
Mean value		0.879	0.889	0.891	0.879
Min. value		0.785	0.786	0.787	0.785
Max. value		0.928	0.931	0.933	0.917

Independent variable: $\hat{\omega}^2$. Dependent variable: net percentage of individual changes based on SID. Sk, skewness; kr, Kurtosis.

DISCUSSION

Our first goal in this paper was to determine whether ABC (quantified by d in the single group design or by $\hat{\omega}^2$ in the control group design) is related to IBC (quantified as the percentage of individual changes, or net percentage in the control group design). Our simulations indicate that percentage of changes is related to average-based effect size. In all conditions, and for both designs, the results show that, as average-based effect size increases, so does the percentage of changes.

Within this general goal, we aimed at finding a mathematical function to capture the relation between effect size and percentage of changes. In both designs, the adjusted linear, quadratic and cubic functions showed excellent fit. The logistic function showed good fit in the single group design, and excellent fit in the control group design. Among them, the linear model was the most parsimonious and easiest to interpret, and hence was preferred (Bentler and Mooijaart, 1989; Maxwell and Delaney, 2004; Steele and Douglas, 2006). It showed excellent fit in all conditions even in the least favorable simulated scenarios ($n = 25$): the R^2 values ranged from 0.90 to 0.96 in the single group design (**Table 2**) and from 0.79 to 0.93 in the control group design (**Table 4**).

Finally, we wanted to identify conditions in which the ABC and IBC are related. Our results indicate that such a relation was present in all simulated conditions and for both designs, regardless of the pre and post distributions skewness, and of the pre-post correlation. The fit (R^2) of the linear regression function slightly varied from 0.96, in the most favorable conditions, to 0.90 (single group design), and 0.79 (control group design) in the most adverse. As sample size increases, so does fit: with $n = 100$, R^2 reached 0.98 in the most favorable conditions, and was never below 0.87 in the most adverse.

A very important finding from our study was that, for both designs, *the slope of the regression line was approximately the same in all simulated conditions*. In the single group design (with d as predictor and the percentage of changes as dependent variable), the slope value was around 30 (ranging from 29 to 31). This indicates that, for each added point to the effect size, the function's estimation of the percentage of changes increased by 30 points. In other words, a 0.10-point increase in d (pre-post differences metric) was associated with a 3-point increase in the percentage of individual changes[4].

In the control group design (with $\hat{\omega}^2$ as predictor and the *net* percentage of changes as the dependent variable), the slope value was around 153 points, ranging from 140 to 165. Because the values of $\hat{\omega}^2$ range from 0 to 1, expressing it this way is more useful: for each 0.10 added points to the effect size, the linear function estimate for the percentage of individual changes increases in 15.3 points (ranging from 14 to 16.5).

Relevance of the Present Findings

Some important implications are worth noting: (a) The ABC and IBC statistics are nearly equivalent; and (b) Cutoffs commonly used for deciding when an effect is small, medium or large should be replaced with more informative indices. Below, we expand on these ideas and offer two recommendations based on them.

The ABC and IBC Statistics Are Nearly Equivalent

With two exceptions (Norman et al., 2001; Lemieux et al., 2007), papers on this topic agree on the following idea: researchers will arrive to different conclusions about a treatment's effectiveness depending on whether they assess it at the individual or at the group level (e.g., Ottenbacher et al., 1988; Testa, 2000; Schmitt and Di Fabio, 2004; Vindras et al., 2012). Our results indicate that this idea is incorrect. Across all of our simulation conditions, ABC and IBC statistics were so closely related that can be considered as different expressions of nearly the same information. This is to be expected, indeed, when variability of pre- and post-test scores is the same. Because increases in effect size lead to increases in the center of the pre-post differences

[4] For a correct interpretation of these results, it should be noted that if the Y variable ranges from zero to 100 (as percentage of changes does) and the X-Y relation is perfect, the Y slope value equals 100 divided by the X range. In our case, if the relationship between X (effect size) and Y (percentage of changes) were perfect, the slope of the regression line will be equal to 100/3.6 = 27.8. The slopes found in this study ranged from 29 to 31 because the studied relationships were not perfect. This only means that, in order to find the correct slope, it is important to take into consideration a range of X values which allows working with all possible values of Y. Our results show that the chosen range of effect size values allowed us to study the complete range of percentages of individual changes

TABLE 5 | Coefficients (and standard errors) for the lineal regression model in the design with a control group.

	$\rho_{\text{pre-post}} = 0.5$		$\rho_{\text{pre-post}} = 0.7$		$\rho_{\text{pre-post}} = 0.9$	
	B_0	B_1	B_0	B_1	B_0	B_1
Sk = −3, Kr = 18	3.51 (0.29)	162.9 (0.82)	4.18 (0.29)	164.7 (0.83)	5.92 (0.31)	164.9 (0.89)
Sk = −2, Kr = 9	3.02 (0.24)	160.3 (0.69)	3.95 (0.25)	162.0 (0.70)	5.18 (0.25)	162.5 (0.72)
Sk = −1, Kr = 2	3.49 (0.21)	156.7 (0.59)	3.74 (0.21)	157.5 (0.59)	4.50 (0.21)	158.2 (0.59)
Sk = 0, Kr = 0	2.35 (0.19)	152.2 (0.53)	2.56 (0.19)	151.7 (0.53)	2.54 (0.19)	151.8 (0.54)
Sk = 1, Kr = 2	1.07 (0.21)	149.8 (0.60)	0.86 (0.22)	147.6 (0.61)	0.37 (0.22)	146.6 (0.61)
Sk = 2, Kr = 9	0.96 (0.26)	148.8 (0.73)	0.33 (0.26)	146.2 (0.72)	−0.03 (0.28)	143.4 (0.77)
Sk = 3, Kr = 18	0.95 (0.30)	147.6 (0.84)	0.63 (0.31)	144.1 (0.85)	0.43 (0.33)	139.7 (0.91)

n = 25; independent variable: $\hat{\omega}^2$; dependent variable: net percentage of individual changes based on SID. Sk, skewness; Kr, kurtosis.

distribution, the number of cases on the right side of any chosen cutoff will also increase.

Based on this finding, we offer our first **recommendation:** When evaluating the change in a group, if only one pre- and one post- measures are available, a logical sequence of analytic steps is as following: (a) assess individual changes through *SID* or *RCI*, (b) aggregate the individual results into a percentage of reliable individual changes (or net percentage, if more than one group is analyzed), and, (c) report this individual-based statistics *along with* classical average-based effect size estimations such as *d* or $\hat{\omega}^2$.

This procedure has several advantages over just reporting the ABC statistics. First, it allows researchers to make decisions about each particular case. This is a common concern in applied settings, and the individual-based methods discussed here provide a straightforward tool for addressing it (Sijtsma, 2012). The usefulness and convenience of these indices have been discussed elsewhere (Jacobson and Truax, 1991; Maassen, 2000; Ferrer and Pardo, 2014). Second, an effect size expressed as a percentage is easier to understand and it enhances the communication of results, especially among researchers without a strong statistical background. For example, in a randomized controlled trial, stating that the effect size was $\hat{\omega}^2 = 0.20$ is less clear than stating that the observed net percentage of individual changes was 33%.

Recent recommendations advocate that effect size estimates should directly address the research question which motivated their estimation, and should be intuitively accessible so that they facilitate the constructive scrutiny of results (Pek and Flora, 2018). We argue that, when used for effect size interpretation, individual-based statistics accomplish both aims. Based on this, and in line with previous work (e.g., Ogles et al., 2001; Wise, 2004; Lambert and Ogles, 2009; Speelman and McGann, 2013; de Beurs et al., 2016; Fisher et al., 2018), we encourage other researchers to include individual-based statistics in their methodological toolbox and to use them to report their results.

Another finding worth highlighting is that, because the intercept and slope coefficients were very similar across conditions, it is easy to compute an approximate percentage (or net percentage) of reliable individual changes *even without having access to the raw data*. For example, if a researcher wants to express an already published effect size as a percentage of changes, the only needed step is to introduce the estimate into the linear regression equation proposed in our results. For example, in a single group pre-post study with $d = 0.9$ with normally distributed scores, and based on **Table 3**:

$$\textit{Percentage of changes} \approx B_0 + B_1 \times d = 1.9 + 29 \times 0.9 \approx 28\% \quad (3)$$

In a control group pre-post study with $\hat{\omega}^2 = 0.4$, with normally distributed scores, pre-post correlation of $r = 0.7$, and based on **Table 5**:

$$\textit{Net percentage of changes} \approx B_0 + B_1 \times \hat{\omega}^2 = 2.6 + 152 \times 0.4 \approx 63\% \quad (4)$$

When *d* or $\hat{\omega}^2$ are not available in the published report, it is easy to compute them from other effect sizes estimates (see Appendix 1 in **Supplementary Data Sheet 2** for examples of these computations, and see Appendix 2 in **Supplementary Data Sheet 2** for an application to data from one published paper). The specific intercept and slope values can be selected according to the empirical skewness and kurtosis (see **Tables 3**, **5**). But even if coefficients from a wrong condition are selected, the estimate of the (net) percentage of changes will be close to the real value.

Based on previous research (Blanca et al., 2013), less than 5% of real datasets have more extreme distributions than the ones simulated here. Consequently, our simple linear regression models can be applied in most real situations to estimate the approximate percentage of individuals who experienced change, when only average-based change indicators are available. Of course, when possible, computing the actual empirical value is preferable.

Cutoffs Commonly Used for Deciding When an Effect Is Small, Medium or Large Should be Replaced With More Informative Indices

In many contexts, it is frequent to use cutoffs to interpret the magnitude of an effect. The cutoffs proposed by Cohen (1988) are arguably the most popular. When considering these cutoffs for identifying small, medium and large effect sizes, we find that, in our simulated single group pre-post scenarios, a small effect ($d = 0.2$) corresponds to 8% of changes, a medium effect ($d = 0.5$) corresponds to 17%, and a large effect ($d = 0.8$) corresponds to

26%. Similar guidelines have been proposed for control group pre-post designs (e.g., Kirk, 2013). According to our results, the proposed $\hat{\omega}^2$ values for declaring a small, moderate, and large effect size (0.01, 0.06, and 0.14) would lead to 4, 12, and 24% of net percentage of changes, respectively. In both designs, the idea that a so-called large effect size leads to just 24–26% of changes (or net changes) does not seem reasonable.

Based on our findings, we recommend that arbitrary cutoffs for evaluating the magnitude of effect estimates should not be used. We are not proposing a new set of cutoffs; rather, we propose to stop using them altogether. Indeed, other authors have suggested this idea before (e.g., Hill et al., 2008; Pek and Flora, 2018), but researchers still use arbitrary guidelines and cutoffs because they are useful for making sense of their findings. Particularly in clinical, educational, and other substantive domains, applied practitioners need to *know the meaning* of values such as $d = 0.6$, $r = 0.4$, $\hat{\eta}^2 = 0.4$, or $\hat{\omega}^2 = 0.35$. Arbitrary cutoffs are appealing as easy rules of thumb, despite their many disadvantages.

Our second recommendation is to use individual-based statistics as a simple tool for interpreting the magnitude of empirical effects. We illustrate this idea with a simple example. Suppose a researcher wants to assess the effectiveness of a new treatment for the pathological fear of darkness. A sample of 100 patients with this fear is gathered and randomly assigned to two groups (treatment group, receiving the new intervention, and control group, receiving no intervention). After finishing the program, the researcher obtains an average-based effect size of $\hat{\omega}^2 = 0.26$ for the interaction between group and occasion of measurement. Instead of declaring that the effect is "large" (Kirk, 2013), the researcher also computes a net percentage of changes (based on **Table 5**),

$$\textit{Net percentage of changes} \approx B_0 + B_1 \times \hat{\omega}^2 = 2.6 + 152 \times 0.26 \approx 42\%.$$

Using the individual-based statistic and substantive knowledge on the disorder, he decides to discard the new intervention in favor of the traditional one, because they usually achieve much higher rates of success. Now, suppose that a different researcher wants to assess the effectiveness of a new treatment for autism in 10- year old children. She applies the new intervention using the exact same sample size and research design, and finds the same effect sizes estimates. In the context of an intervention to treat autism spectrum disorders, she can arguably claim that the effect is "very large" (indeed, she can claim the Nobel Prize).

In both cases, the researchers can easily decide whether $\hat{\omega}^2 = 0.26$ means a "small" or a "large" effect based on: (a) the individual-based statistic; and (b) their theoretical knowledge on the substantive domain. The individual-based statistics help interpreting the meaning of the effect size estimation but, unlike arbitrary "general guidelines," do not force researchers to interpret them invariantly across different domains. By using them, applied practitioners can easily understand and communicate the meaning of any value of the percentage of changes in the context of their particular field.

Theoretical and Methodological Considerations, and Future Directions

In our analyses we used the standard deviation of the pre-post differences (σ_{dif}) as the standardizer of our single-group ABC statistic, but other standardizers are also available. For example, one common procedure is to use the standard deviation of the pre- scores (σ_{pre}). The choice of the standardizer is related to the ability of the effect size measure to deal with pre-post dependency. Using σ_{dif} allows taking into account such dependency because σ_{dif} is partially dependent on the pre-post correlation, but there is no consensus on the correct procedure, and different authors advocate for different solutions (Gibbons et al., 1993; Dunlap et al., 1996; Morris and DeShon, 2002; Ahn et al., 2012).

A full discussion of the implications of using different standardizers is beyond the scope of this study, and we refer the reader to the aforementioned literature. However, it is important to note that using σ_{pre} as the standardizer for d will affect the relation between the ABC and IBC statistics. Specifically, the B_1 coefficient in Equation 3, which captures the regression slope, will have higher values for higher levels of pre-post correlation. In other words, although the relation can be considered linear regardless of the standardizer chosen, the slope of such linear function will differ depending on the pre-post correlation if σ_{pre} is used. In contrast, it will remain constant if σ_{dif} is chosen. See Appendix 1 in **Supplementary Data Sheet 2** for a more detailed description and some examples.

In a different vein, some caution is warranted when interpreting IBC statistics. For example, suppose that a researcher assesses the change in academic achievement from a given grade to the next in a single school group, and she finds an effect size of $d = 0.3$. With normally distributed scores, and according to our Equation 3,

$$\textit{Percentage of changes} \approx B_0 + B_1 \times d = 1.9 + 29 \times 0.3 \approx 11\%$$

This value does not imply that 89% of the students did not learn. Instead, it indicates that, given the observed variability in the pre-post differences, only 11% of such improvements could be identified as reliable. The same mean difference (say, for example, 10 IQ points) combined with a lower value of σ_{dif} would lead to a higher value of both d and the percentage of changes. Note that this "attenuation problem" affects both the ABC and IBC statistics. Other factors such as measurement error also attenuate the value of both types of statistics (see the Appendix 3 in **Supplementary Data Sheet 2**). However, IBC statistics should always be interpreted in the context of a particular research domain, and it is reasonable to think that measurement error, "natural" variability in the differences (σ_{dif}), and other attenuating factors, will remain fairly constant across studies from the same domain—particularly if they use the same measurement instrument. If more than two measurement occasions are available, other statistical tools can be used to assess individual change (e.g., Estrada et al., 2018). These tools are particularly useful for examining developmental and learning processes, and can incorporate measurement error.

In our simulated scenarios, both groups were expected to have scores with the same distributional shape and dispersion in the pre- and post- evaluations—i.e., only the center of the distribution was expected to change. Of course, the distribution shape and variability can also change between both assessments, for example, as a result of an intervention. It is unclear whether our findings apply to such scenarios, and future research should address this important point.

CONCLUSION

In this paper we show that individual- and average-based statistics for measuring change are closely related, regardless of sample size, pre-post correlation, and shape of the scores' distribution. To our knowledge, this is the first study applying individual reliable change indices to an experimental design. Our findings are relevant for a range of scientific disciplines including education, psychology, medical and physical therapy. We encourage other researchers to use individual change indices and individual-based statistics. Their main advantages are: (a) they allow determining which individual cases changed reliably; (b) they facilitate the interpretation and communication of results; and (c) they provide a straightforward evaluation of the magnitude of empirical effects while avoiding the problems of arbitrary general cutoffs.

AUTHOR CONTRIBUTIONS

AP developed the original idea. EE and AP reviewed the relevant literature. EE and AP designed the study and data analysis strategy. EE conducted the data simulation. EE, EF, and AP analyzed the data and interpreted the results, organized the article structure and drafted the manuscript critically revised the manuscript.

FUNDING

EE was supported by the scholarship FPI-UAM 2011 (granted by Universidad Autónoma de Madrid). The publication fee was partially supported by the Library of UC, Davis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02696/full#supplementary-material

Supplementary Data Sheet 1 | Simulation code.

Supplementary Data Sheet 2 | Appendices.

REFERENCES

Ahn, S., Ames, A. J., and Myers, N. D. (2012). A review of meta-analyses in education: methodological strengths and weaknesses. *Rev. Educ. Res.* 82, 436–476. doi: 10.3102/0034654312458162

Bauer, S., Lambert, M. J., and Nielsen, S. L. (2004). Clinical significance methods: a comparison of statistical techniques. *J. Pers. Assess.* 82, 60–70. doi: 10.1207/s15327752jpa8201_11

Beaton, D. E., Bombardier, C., Katz, J. N., and Wright, J. G. (2001). A taxonomy for responsiveness. *J. Clin. Epidemiol.* 54, 1204–1217. doi: 10.1016/S0895-4356(01)00407-3

Bentler, P. M., and Mooijaart, A. (1989). Choice of structural model via parsimony: a rationale based on precision. *Psychol. Bull.* 106, 315–317. doi: 10.1037/0033-2909.106.2.315

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodol. Eur. J. Res. Methods Behav. Soc. Sci.* 9, 78–84. doi: 10.1027/1614-2241/a000057

Christensen, L., and Mendoza, J. L. (1986). A method of assessing change in a single subject: an alteration of the RC index. *Behav. Ther.* 17, 305–308. doi: 10.1016/S0005-7894(86)80060-0

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: L. Erlbaum Associates.

Crawford, J. R., and Howell, D. C. (1998). Regression equations in clinical neuropsychology: an evaluation of statistical methods for comparing predicted and obtained scores. *J. Clin. Exp. Neuropsychol.* 20, 755–762. doi: 10.1076/jcen.20.5.755.1132

Crawford, J. R., Howell, D. C., and Garthwaite, P. H. (1998). Payne and jones revisited: estimating the abnormality of test score differences using a modified paired samples t test. *J. Clin. Exp. Neuropsychol.* 20, 898–905. doi: 10.1076/jcen.20.6.898.1112

Crosby, R. D., Kolotkin, R. L., and Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *J. Clin. Epidemiol.* 56, 395–407. doi: 10.1016/S0895-4356(03)00044-1

Cumming, G., and Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educ. Psychol. Meas.* 61, 532–574. doi: 10.1177/0013164401614002

de Beurs, E., Barendregt, M., de Heer, A., van Duijn, E., Goeree, B., Kloos, M., et al. (2016). Comparing methods to denote treatment outcome in clinical research and benchmarking mental health care. *Clin. Psychol. Psychother.* 23, 308–318. doi: 10.1002/cpp.1954

de Vet, H. C., Ostelo, R. W., Terwee, C. B., van der Roer, N., Knol, D. L., Beckerman, H., et al. (2006). Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual. Life Res.* 16, 131–142. doi: 10.1007/s11136-006-9109-9

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., and Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol. Methods* 1, 170–177. doi: 10.1037/1082-989X.1.2.170

Estrada, E., Ferrer, E., Shaywitz, B. A., Holahan, J. M., and Shaywitz, S. E. (2018). Identifying atypical change at the individual level from childhood to adolescence. *Dev. Psychol.* 54, 2193–2206. doi: 10.1037/dev0000583

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychol. Methods* 14, 43–53. doi: 10.1037/a0014699

Ferrer, R., and Pardo, A. (2014). Clinically meaningful change: false positives in the estimation of individual change. *Psychol. Assess.* 26, 370–383. doi: 10.1037/a0035419

Fisher, A. J., Medaglia, J. D., and Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl. Acad. Sci. U.S.A.* 115, E6106–E6115. doi: 10.1073/pnas.1711978115

Fisher, R. A. (1935). The logic of inductive inference. *J. R. Stat. Soc.* 98, 39–82. doi: 10.2307/2342435

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol. Gen.* 141, 2–18. doi: 10.1037/a0024338

Gibbons, R. D., Hedeker, D. R., and Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *J. Educ. Stat.* 18, 271–279. doi: 10.3102/10769986018003271

Grissom, R. J., and Kim, J. J. (2012). *Effect Sizes for Research: Univariate and Multivariate Applications, 2nd Edn.* New York, NY: Routledge.

Guyatt, G. H., Juniper, E. F., Walter, S. D., Griffith, L. E., and Goldstein, R. S. (1998). Interpreting treatment effects in randomised trials. *BMJ* 316, 690–693. doi: 10.1136/bmj.316.7132.690

Hageman, W. J., and Arrindell, W. A. (1999). Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis. *Behav. Res. Ther.* 37, 1169–1193.

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

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Dev. Perspect.* 2, 172–177. doi: 10.1111/j.1750-8606.2008.00061.x

Jacobson, N. S., Follette, W. C., and Revenstorf, D. (1984). Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behav. Ther.* 15, 336–352. doi: 10.1016/S0005-7894(84)80002-7

Jacobson, N. S., Roberts, L. J., Berns, S. B., and McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *J. Consult. Clin. Psychol.* 67, 300–307. doi: 10.1037/0022-006X.67.3.300

Jacobson, N. S., and Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59, 12–19. doi: 10.1037/0022-006X.59.1.12

Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6829–6833. doi: 10.1073/pnas.0801268105

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

Kirk, R. E. (2013). *Experimental Design: Procedures for the Behavioral Sciences, 4th Edn*. Thousand Oaks, CA: Sage Publications.

Lambert, M. J., and Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: the need for a common procedure and validity data. *Psychother. Res.* 19, 493–501. doi: 10.1080/10503300902849483

Lemieux, J., Beaton, D. E., Hogg-Johnson, S., Bordeleau, L. J., and Goodwin, P. J. (2007). Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. *J. Clin. Epidemiol.* 60, 448–455. doi: 10.1016/j.jclinepi.2006.08.006

Lowrie, T., Logan, T., and Ramful, A. (2017). Visuospatial training improves elementary students' mathematics performance. *Br. J. Edu. Psychol.* 87, 170–186. doi: 10.1111/bjep.12142

Lydick, E., and Epstein, R. S. (1993). Interpretation of quality of life changes. *Qual. Life Res.* 2, 221–226. doi: 10.1007/BF00435226

Maassen, G. H. (2000). Kelley's formula as a basis for the assessment of reliable change. *Psychometrika* 65, 187–197. doi: 10.1007/BF02294373

Maassen, G. H. (2004). The standard error in the jacobson and truax reliable change index: the classical approach to the assessment of reliable change. *J. Int. Neuropsychol. Soc.* 10, 888–893. doi: 10.1017/S1355617704106097

Maxwell, S. E., and Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, A., Dietzfelbinger, L., Rosseel, Y., and Steyer, R. (2016). The effectliter approach for analyzing average and conditional effects. *Multivariate Behav. Res.* 51, 374–391. doi: 10.1080/00273171.2016.1151334

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105, 156–166. doi: 10.1037/0033-2909.105.1.156

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organ. Res. Methods* 11, 364–386. doi: 10.1177/1094428106291059

Morris, S. B., and DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol. Methods* 7, 105–125. doi: 10.1037/1082-989X.7.1.105

Muroff, J., Steketee, G., Frost, R. O., and Tolin, D. F. (2014). Cognitive behavior therapy for hoarding disorder: follow-up findings and predictors of outcome. *Depress. Anxiety* 31, 964–971. doi: 10.1002/da.22222

Norman, G. R., Sridhar, F. G., Guyatt, G. H., and Walter, S. D. (2001). Relation of distribution and anchor-based approaches in interpretation of changes in health-related quality of life. *Med. Care* 39, 1039–1047. doi: 10.1097/00005650-200110000-00002

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory, 3.a Edn*. New York, NY: McGraw-Hill.

Ogles, B. M., Lunnen, K. M., and Bonesteel, K. (2001). Clinical significance: history, application, and current practice. *Clin. Psychol. Rev.* 21, 421–446. doi: 10.1016/S0272-7358(99)00058-6

Ottenbacher, K. J., Johnson, M. B., and Hojem, M. (1988). The significance of clinical change and clinical change of significance: issues and methods. *Am. J. Occup. Ther.* 42, 156–163. doi: 10.5014/ajot.42.3.156

Payne, R. W., and Jones, H. G. (1957). Statistics for the investigation of individual cases. *J. Clinic. Psychol.* 13, 115–121.

Pearl, J. (2009). *Causality*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511803161

Pedhazur, E. J., and Schmelkin, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pek, J., and Flora, D. B. (2018). Reporting effect sizes in original psychological research: a discussion and tutorial. *Psychol. Methods* 23, 208–225. doi: 10.1037/met0000126

Pi-Sunyer, X., Astrup, A., Fujioka, K., Greenway, F., Halpern, A., Krempf, M., et al. (2015). A randomized, controlled trial of 3.0 mg of liraglutide in weight management. *N. Eng. J. Med.* 373, 11–22. doi: 10.1056/NEJMoa 1411892

Revicki, D., Hays, R. D., Cella, D., and Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J. Clin. Epidemiol.* 61, 102–109. doi: 10.1016/j.jclinepi.2007.03.012

Richardson, E. (1996). Measures of effect size. *Behav. Res. Methods Instrum. Comput.* 28, 12–22. doi: 10.3758/BF03203631

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701. doi: 10.1037/h0037350

Schmitt, J. S., and Di Fabio, R. P. (2004). Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J. Clin. Epidemiol.* 57, 1008–1018. doi: 10.1016/j.jclinepi.2004.02.007

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Sijtsma, K. (2012). Future of psychometrics: ask what psychometrics can do for psychology. *Psychometrika* 77, 4–20. doi: 10.1007/s11336-011-9242-4

Speelman, C., and McGann, M. (2013). How mean is the mean? *Front. Psychol.* 4:451. doi: 10.3389/fpsyg.2013.00451

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

Testa, M. A. (2000). Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. *Med. Care 38*, II166–II174. doi: 10.1097/00005650-200009002-00026

Vindras, P., Desmurget, M., and 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:e39059. doi: 10.1371/journal.pone.0039059

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. *J. Pers. Assess.* 82, 50–59. doi: 10.1207/s15327752jpa8201_10

Wyrwich, K. W., Tierney, W. M., and Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J. Clin. Epidemiol.* 52, 861–873.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.