Assigning and Combining Probabilities in Single-Case Studies

R. Manolov[1], A. Solanas[2]

Department of Behavioral Sciences Methods, University of Barcelona, Barcelona, Spain

[1]*rrumenov13@ub.edu,* [2]*antonio.solanas@ub.edu*

Abstract

There is currently a considerable diversity of quantitative measures available for summarizing the results in single-case studies (e.g., nonoverlap indices, regression coefficients or R-squared values). Given that the interpretation of some of them is difficult due to the lack of established benchmarks, the current paper proposes an approach for obtaining further numerical evidence on the importance of the results, complementing substantive criteria, visual inspection and the summary indices themselves. This additional evidence consists in obtaining the likelihood of the outcome in case the intervention was ineffective. The probability is expressed in terms of *p* values, which can then be used to integrate the results of several studies; an integration which is problematic when different metrics are used across primary studies and raw data are not available. Two methods for combining probabilities in the context of single-case studies are pointed out – one based on a weighted average and the other on the binomial test.

Introduction

Currently there is a great variety of techniques proposed for quantifying the magnitude of effect in single-case data. However, not all of these procedures are accompanied by unquestionable interpretative benchmarks in order to judge the relevance of the results obtained. Even in the case of regression-based procedures which yield R-squared values the use of Cohen's guidelines in single-case studies has been put in doubt due to greater effects usually found [1]. Apart from the interpretation of individual studies' results, another important question yet to be answered is how to integrate the results of several studies conducted on the same topic. As regards integration meta-analysis is the option of choice, but combined significance may be useful when effect sizes are not reported or there is not an established effect size measure for some data analytic procedure [2]. Furthermore, the great proliferation of analytical techniques expressed in different metrics and the lack of consensus on which technique to use in order to summarize the results hinders carrying out meta-analyses [3] and opens the possibility for combining probabilities. Note, however, that when raw data are available the researcher can compute the metric of choice regardless of the original primary indicator and combine the results afterwards. The present study focuses on two complementary topics: a) the additional assessment of intervention effectiveness in an individual single-case study – this is achieved estimating statistical significance after constructing the relevant sampling distribution and b) the quantitative integration of several single-case studies using different indices for quantifying the magnitude of effect.

Additional evidence for effectiveness in individual studies

Rationale

Single-case study results should be analyzed both visually and numerically, while also using substantive criteria of what effect is relevant in the specific behavioral context. However, it would also be useful for the researcher to have a statistical criterion complementing the substantive one. Given the lack (or inadequacy) of benchmarks for most indices, we propose a method for further assessing the relevance of an effect size. This additional evidence is based on obtaining the statistical significance associated with the index computed. *p* values have already been used in single-case designs via randomization tests [4], but in the proposal made here the sampling distribution is not constructed after permuting the data or the points of change in phase. The basic idea is that each effect size index, standardized or unstandardized, has its sampling distribution and its expected value in the conditions of no intervention effect, that is, when the null hypothesis is true. The value actually obtained (i.e., the

"outcome") can be located in the sampling distribution in order to know whether a value as large as or larger than the outcome is likely to be obtained only by chance in absence of effect.

The maximal reference approach (MRA)

The current proposal is related to simulation modeling analysis, SMA [5], in which the outcome is located in a sampling distribution based on generated samples extracted from a population where the following parameters are specified: phase lengths equal to the ones of the original sample, no intervention effect, phase serial dependence equal to the autocorrelation estimated in the original sample, normal disturbance. However, at least two assumptions are being made with the SMA: a) data are normal, which may be questionable, and b) the autocorrelation is estimated precisely, which is problematic when few measurements are available. In case these assumptions are not met, the sampling distribution constructed may not be appropriate. In order to reduce the uncertainty around these unknown data features and gain confidence on the validity of the sampling distribution as a reference, it may be necessary to construct not one but several sampling distributions, provided that the typical values of the primary indicators are expected to vary according to series/phase lengths, the data generation processes, the degrees of serial dependence, the random variable distributions, etc. Our proposal is to follow a conservative approach, the maximal reference approach (MRA), in which the index values associated with several key *p* values (e.g., .90, .80, ..., .20, .10, .05) are identified for several conditions. If the *p* values are tabulated, then a researcher can compare the outcome to the reference values from the table in order to know in what range of probability is such an outcome expected at random. For instance, suppose that a researcher obtains an R-squared value of .60 after carrying out a regression analysis, whereas another researcher obtains a value of 95% using a nonoverlap index. Are the effects large? In which study is the intervention effect stronger? We consider that the MRA, when used jointly with visual analysis and substantive criteria, may aid professionals interested in these questions.

Quantitative integration of single-case studies using different metrics

Combining weighted vs. unweighted probabilities.

One of the possible sources of invalidity when integrating results is weighting equally studies with different sample sizes, but other factors affecting the reliability and validity of the results should also be considered [6]. Weighting gives more information and has proven to be more powerful when combining studies with different sample sizes [7]. Additionally, weighting is already inherent to the meta-analytical combination of effect sizes.

Methods for combining probabilities.

Of the diversity of existing approaches for combining probabilities only two will be highlighted here, given that they are especially relevant for the approach presented in the previous section. One of the methods for combining probabilities that can be used is similar to a proposal consisting in testing the statistical significance of the mean of the *p* values of the studies to be integrated [8]. However, as we underline the need of weighting the *p* values, the statistical test presented by Edgington [8] is not applicable. As regards the exact weighting procedure, a minimal requirement would be to use series length (the single-case equivalent of sample size) as a weight. However, it should be considered that the ideal weight is the inverse of the error variance, that is, the Fisher information [7]. Therefore, one option would be to use the inverse of the variance of the summary index, given that the amount of dispersion of the index values about its mathematical expectancy is closely related to the reliability of the results. It has to be highlighted that this approach is equivalent to the one usually followed in the meta-analytical integration of studies' findings. In single-case designs, using the index's variance is impeded by the fact that different studies use different metrics. Given that variances are not directly comparable, we propose using the coefficient of variation (CV). In order to follow the MRA, for each outcome the researcher should assign a probability and a CV, both which need to be made available, for instance, in a tabular format. Given that the exact *p* value associated with the outcome is not known in the MRA, Edgington's [9] conservative solution can be followed, taking the upper limit of the probability given by a table as the probability itself (e.g., if the information available is that $p < .05$, then $p = .05$ should be the value used when integrating studies). The

weights obtained via the CV could then be used to compute a weighted mean of the *p* values which can be compared to a predefined reference value.

The other method for combining probabilities potentially used for single-case studies is the binomial test, which has been deemed both quick and simple [10]. The main advantage is that it allows using studies which report only information on whether the *p* value was above or below .05 (i.e., it is possible to follow the MRA). The test consists in comparing each *p* value associated with the outcome to the nominal level α predetermined by the researcher. All outcome *p* values lower than α are counted as "successes" (versus "failures" in case of greater values) in terms commonly used when working with the binomial distribution [10]. After that the probability of obtaining as many successes, denoted by *s*, is referred to a binomial distribution with *n* (the number of trials), which is equal to the number of independent individual studies, and π (the probability of success in each trial), which is equal to α. The probability of obtaining *s* or more successes can be calculated directly from $\sum_{x=s}^{n}\binom{n}{x}\alpha^{x}\left(1-\alpha\right)^{n-x}$. For instance, suppose that a researcher wants to integrate the results of ten studies using the same intervention and similar participants, but not necessarily the same effect size indicator. Further suppose that the α selected is .05 and that three out of the ten studies are assigned $p < .05$ via the MRA. In this case, the probability of obtaining such a result only by chance is .0115, which indicates that, when considered together, the studies point at a strong intervention, although such an interpretation is always conditional on the professional's criteria, client's perceptions, etc.

Discussion

Strengths and limitations of the additional evidence for effectiveness in individual studies

The additional evidence has been presented in terms of *p* values, that is, the probability of obtaining as large as or larger outcome only by chance. Nonetheless, this evidence can also be expressed as the proportion of outcomes lower than or equal to the outcome, an idea similar to the one underlying percentiles. In any case, given the limitations of *p* values, these should not be used as the sole indicator, but rather as a complement to practitioner's experience, visual analysis, and the index quantifying behavioral change. The MRA is based on the idea of creating several plausible scenarios which allows researchers to make solidly supported decision. In terms of the efforts required from the researcher, when using the MRA it is only necessary to make a comparison to an already available maximal reference for the index and phase lengths used. However, given that the greatest *p* value of all conditions is used, this approach is rather conservative, which can be thought of as loss of power. The loss of power is attenuated for those cases in which there is evidence on the data features in a specific field and there is a narrower set of scenarios that need to be represented by the sampling distributions. Moreover, the probability assigned will be closer to the actual one – and the approach will be less conservative – in case there is a greater amount of reference values, for instance, for *p* values of .90, .80, …, .10, .05, and .01. Finally, it is not possible to rule out the possibility of publication bias – researchers may feel more inclined to submit for publication results for which the *p* value associated with the outcome is of specific magnitude. Nonetheless, journal editors endorse full reporting and maximal information around the effect size computed and they are not expected to reject an article only on the basis of one indicator, the p value, which is not the main focus of the findings.

Strengths and limitations of the quantitative integration of single-case studies using different metrics

The MRA allows combining studies whose results are summarized using different indices, which is important, given that the lack of a common effect size in single-case designs. The present paper focuses specifically on two ways of combining individual studies' *p* values. The weighted means approach has the drawback of the impossibility to test the composite *p* statistically. It only allows a comparison with a reference *p* determined by the researcher prior to carrying out the combination of individual studies' results. The binomial test is based on counting the amount of studies in which the *p* values is below a predefined nominal α and seems the most natural complement of the MRA. Another advantage of this method is that its logic is relatively simple to understand and its use is straightforward. The binomial test allows specifying any nominal α, given that it uses this reference

as the probability of a positive result in for each study to integrate. Additionally, although the test is performed on *p* values, the original information expressed in terms of the magnitude of effect indices commonly used in single-case research is not lost. In case the idea subjacent to this approach for obtaining additional evidence and integrating it quantitatively is accepted by applied researchers, a logical step would be to obtain the reference values for the most frequently used procedures and for the variety of data patterns described here.

References

1. Parker, R.I., Brossart, D.F., Vannest, K. J., Long, J.R., Garcia De-Alba, R., Baugh, F.G., Sullivan, J.R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review* **34**, 116-132.

2. Becker, B.J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin* **102***,* 164-171.

3. Beretvas, S.N., Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention* **2**, 129-141.

4. Edgington, E.S., Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall/CRC.

5. Borckardt, J.J., Nash, M.R., Murphy, M.D., Moore, M., Shaw, D., O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist* **63**, 77-95.

6. Noble, J.H. (2006). Meta-analysis: Methods, strengths, weaknesses, and political uses. *Journal of Laboratory and Clinical Medicine* **147**, 7-20.

7. Whitlock, M.C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* **18**, 1368-1373.

8. Edgington, E.S. (1972b). A normal curve method for combining probability values from independent experiments. *Journal of Psychology* **82**, 85-89.

9. Edgington, E.S. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology* **80**, 351-363.

10. Darlington, R.B., Hayes, A.F. (2000). Combining independent p values: Extensions of the Stouffer and binomial methods. *Psychological Methods* **5**, 496-515.