Is Busk and Serlin's measure of therapy effect size *d* a suitable measure for use in therapy studies? Evidence from simulations.

Introduction

Beeson & Robey (2006) suggest that Busk and Serlin's (1992) d_1 (henceforth d_{BS}) is the best measure of effect size, and one that should be routinely reported in therapy studies and is suitable for use in meta-analysis of single case therapy studies.

Our impression is that this measure is widely used and reported in single case studies

What are the criteria for a good of effect size suitable for meta-analysis?

We would suggest the following:

1. The measure of effect size should be unbiased in the standard statistical sense.
2. *p* values and confidence intervals can be calculated from the effect size.
3. The effect size measure should be directly related to the amount of improvement.
4. More rapid improvement should be directly reflected in a larger effect size.
5. It should be sensitive to trend across a set of baseline trials.

We investigated the properties of d_{BS} in a large number of simulations. We were particularly interested in whether it was affected by *autocorrelation* – the tendency for performance on one session to be related to performance on previous sessions. There are two possible aspects to this.

The first is *session-level dependence*; the overall probability of correct performance can vary depending on the level of performance in the previous session. The second, independent factor is *item-level dependence*. If an item is correct on one occasion it will be more likely to named correctly the next time it is presented than if it had been incorrectly named on the previous occasion.

Either aspect will result in autocorrelation that is well known to threaten statistical analysis of time series data.

Method

We conducted a set of simulations of patient data varying the following factors:

(i) *auto*: this is the lag 1 autocorrelation between the underlying probability correct on trial *n* and trial *n+1*. This was varied from 0 to 0.25, 0.5 and 0.75. *auto* represents *session-level dependence*.

(ii) *k*: this is the odds ratio for correct performance on trial *n* for items correct on trial n-1 relative to items incorrect on trial n-1. This was varied from 1 to 5, 20 and 100. The value of *k* represents *item-level dependence*.
Together *auto* and *k* result in a measured degree of lag 1 autocorrelation that we call *lag1r*.

(iii) *n*: this is the number of items in the test. This was varied from, 10 to 20, 50, 100.

(iv) *trials*: this is the length of the baseline sequence. This was varied from 3 to 5, 10, 15 and 20.

(v) *intendedsd*: this is the target sd for the sd of the underlying probability correct. It represents the degree of session-to-session variability related to session-level dependence. This was varied from 0.10 to 0.05, 0.025 and 0.01. The obtained sd was necessarily different and is called σ_r .

In addition the model was run with *auto*=0 and *intendedsd=0*, varying *k, n* and *trials* as before.

In each case we generated a sequence of 10020 trials. We defined the *true* d_{BS} as *(0.90-mean baseline)/sd baseline*. (The 0.9 is arbitrary and unimportant). We compared the *measured* d_{BS} in each simulation from *n* items over *trials* trials. Across the 1360 combinations of conditions, each involving 10000 simulations, we could investigate, using simultaneous multiple regression how sample estimates of *dBS* are related to *true dBS* and how d_{BS} estimates are affected by *n*, *trials (t)*, *logk,* σ_r*, lag1r* and *auto*.

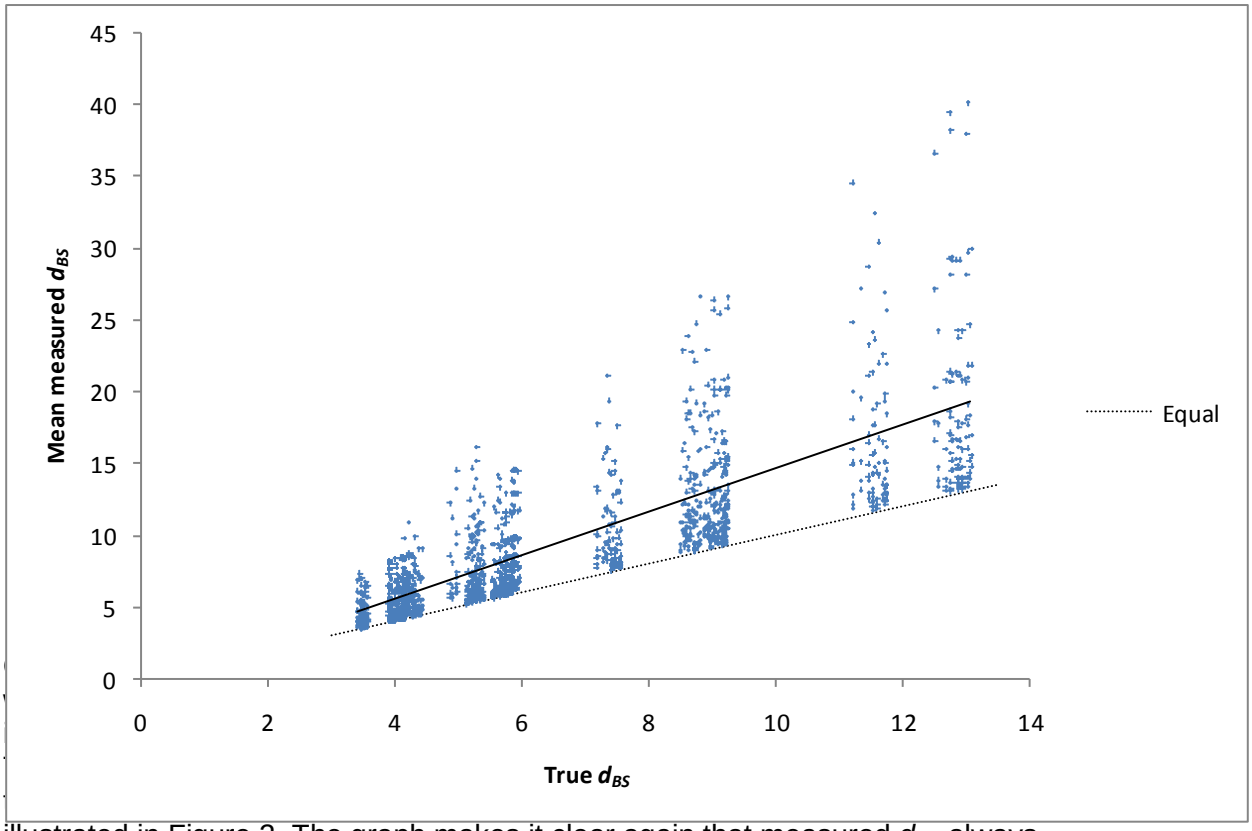
Results

On average 0.9% of runs had no variation in the sampled baseline (with a range from 0 to 25.6%) resulting in d_{BS} of infinity. These runs are eliminated in the results that follow.

In *every* one of the 1360 simulations the mean d_{BS} was significantly greater than the *true* d_{BS} (t test p<.001; see Figure 1).
The overestimation varies from 2.3% to 211% with a mean of 41.3%. Recall that *mean* d_{BS}, is always calculated over 10000 observations; individual values of d_{BS} will, of course, be much more variable.
In every case (of the 1360 simulations), *mean* d_{BS} is significantly greater than the *true* d_{BS}. Clearly, it is a biased estimator of the *true* d_{BS} effect size.

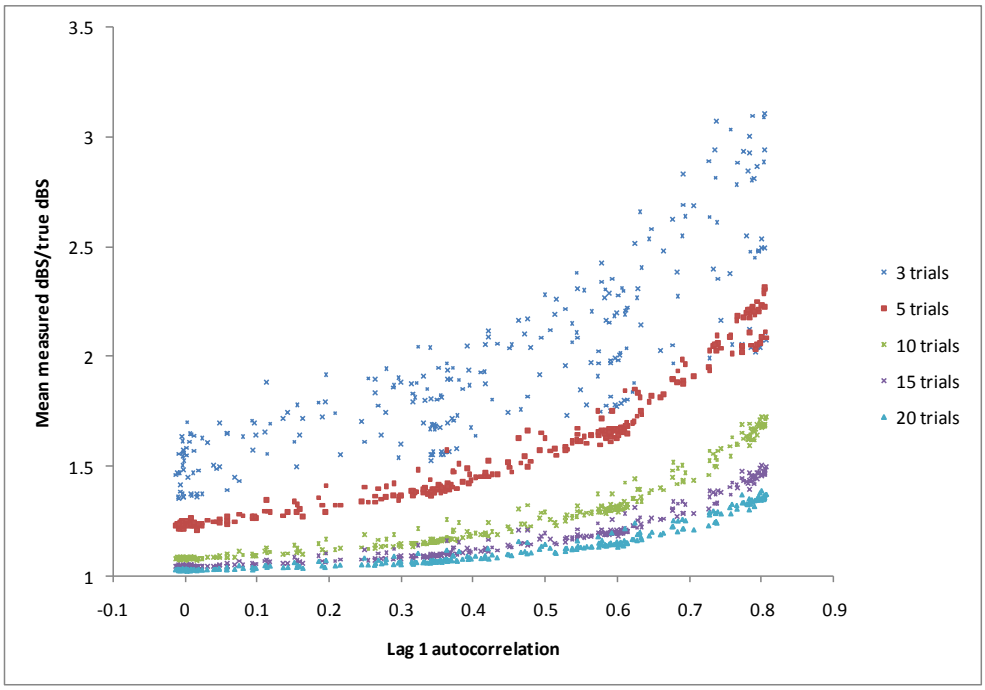
The determinants of overestimation of true d_{BS}.
We investigated how the degree of overestimation (defined as *mean* d_{BS}/*true* d_{BS}) was related to *n, auto, logk, t,* σ_r, and *lag1r* using simultaneous multiple regression.
The degree of overestimation was related to *t* (t(1353)=48.70, p<.001, η^2 = 0.64), *lag1r* (t(1353)=12.06, p<.001, η^2 = 0.097), *n* (t(1353)=5.16, p<.001, η^2 = 0.019) and *logk* (t(1353)= 3.78, p<.001, η^2 = 0.010), but not *auto* (t(1353)=1.62, p=.11, η^2 = 0.002), or σ_r (t(1353)=0.97, p=.33, η^2 = 0.001).

Figure 1. The relationship between *mean* d_{BS} and *true* d_{BS}. The dotted line is represents equal values of mean measured d_{BS} and true d_{BS}. The solid line is the best-fitting line for the 1360 observations.



illustrated in Figure 2. The graph makes it clear again that measured d_{BS} always overestimates the *true* d_{BS}; this is true even where there is no lag 1 autocorrelation. The degree of overestimation is always worse when there are fewer trials in the baseline.

Figure 2. The relationship between mean overestimation of d_{BS}, the number of trials in the baseline and *lag1r*: the autocorrelation in the baseline.



Discussion

So, as a result of these simulations, what do we know about how d_{BS} behaves in relation to the criteria we advanced in the introduction?

(i) *dBS is a biased measure.*
In the simulations, d_{BS} is widely variable but its mean is, in every case, greater than the true value. The degree to which it overestimates the true d_{BS} is primarily related to two factors: it is greater with fewer trials in the baseline and with more autocorrelation in the baseline series (other factors have significant, though substantially smaller, effects on mean d_{BS}).

(ii) *The effect size has a direct interpretation.*
It is clear that d_{BS} varies with a number of factors, but it is not clear how its value is to be interpreted.

(iii) *p values and confidence intervals can be calculated from the effect size.*
We know of no way that *p* values (and hence confidence intervals) can be calculated from d_{BS}. The result is that readers cannot easily discriminate between therapy effect sizes that might easily have occurred by chance and those that are a result of real improvement.

(iv) *The effect size measure should be directly related to the amount of improvement.*
In data not presented here we show that the absolute amount of improvement and d_{BS} is not linear.

(v) *More rapid improvement should be directly reflected in a larger effect size.*
d_{BS} as a measure takes no account of the number of sessions in therapy. As a result, it is clearly unable to capture anything about the rate of improvement.

(vi) *It should be sensitive to trend across a set of baseline trials.*
d_{BS} takes no account of any trend across the baseline trials.

To summarise, d_{BS} as a measure of effect size does not meet *any* of the criteria we suggested as necessary for an effect size measure suitable for meta-analysis that we think are uncontroversial.

References

Beeson, P. M., & Robey, R. R. (2006). Evaluating Single-Subject Treatment Research: Lessons Learned from the Aphasia Literature. *Neuropsychology Review, 16*(4), 161-169.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-Case Research Design and Analysis: New directions for psychology and education.* Hillsdale, N.J.: Lawrence Erlbaum Associates.