PSICOSTAT

IT STARTED!
PSICOSTAT 3.1
MEETINGS' CALENDAR

Design analysis and mixed-effects models

Suggestions on testing treatment efficacy with small effects and small samples

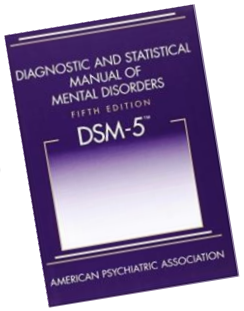
Enrico Toffalini
enrico.toffalini@unipd.it

with the collaboration of David Giofrè, Massimiliano Pastore, Federica Fraccadori, Barbara Carretti, & Denes Szucs

December 18, 2020

Dyslexia treatment as an ideal case example

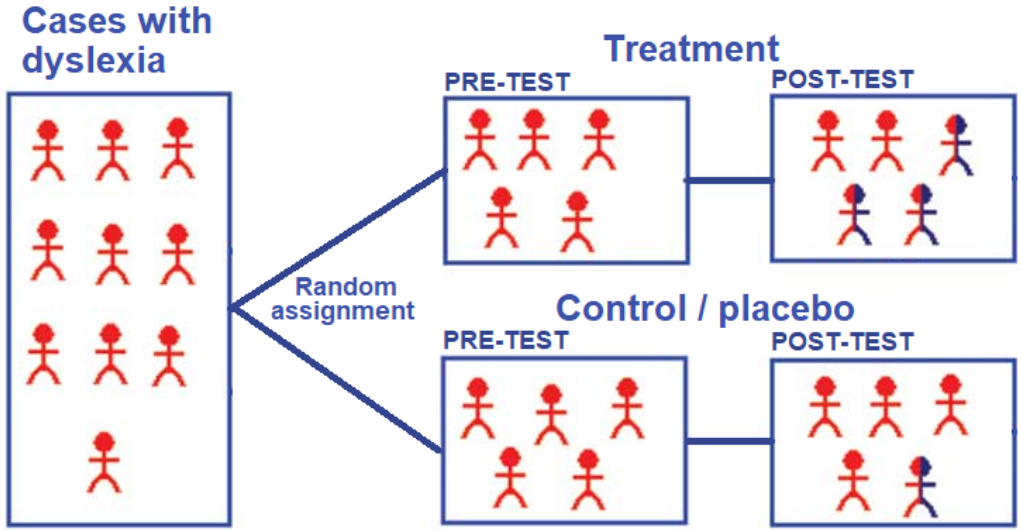
- **The effect size is (almost a priori) small**
  (effects in psychology; DSM-5: resistance to treatment)



- **Collecting large samples is very difficult**
  (relatively small sub-population; much compliance required for the study)

- **Outcome [reading] is easy to measure; measures have good, but not perfect, reliability**

Randomized controlled trial (RCT)



individual outcomes are quantitative, not binomial

Data analysis *(Dimitrov & Rumrill, 2003; Gelman, 2020; Van Breukelen, 2006)*

- ANCOVA / LM on post-test with pre-test covaried

lm(score[time==«post»] ~ **group** + score[time==«pre»])

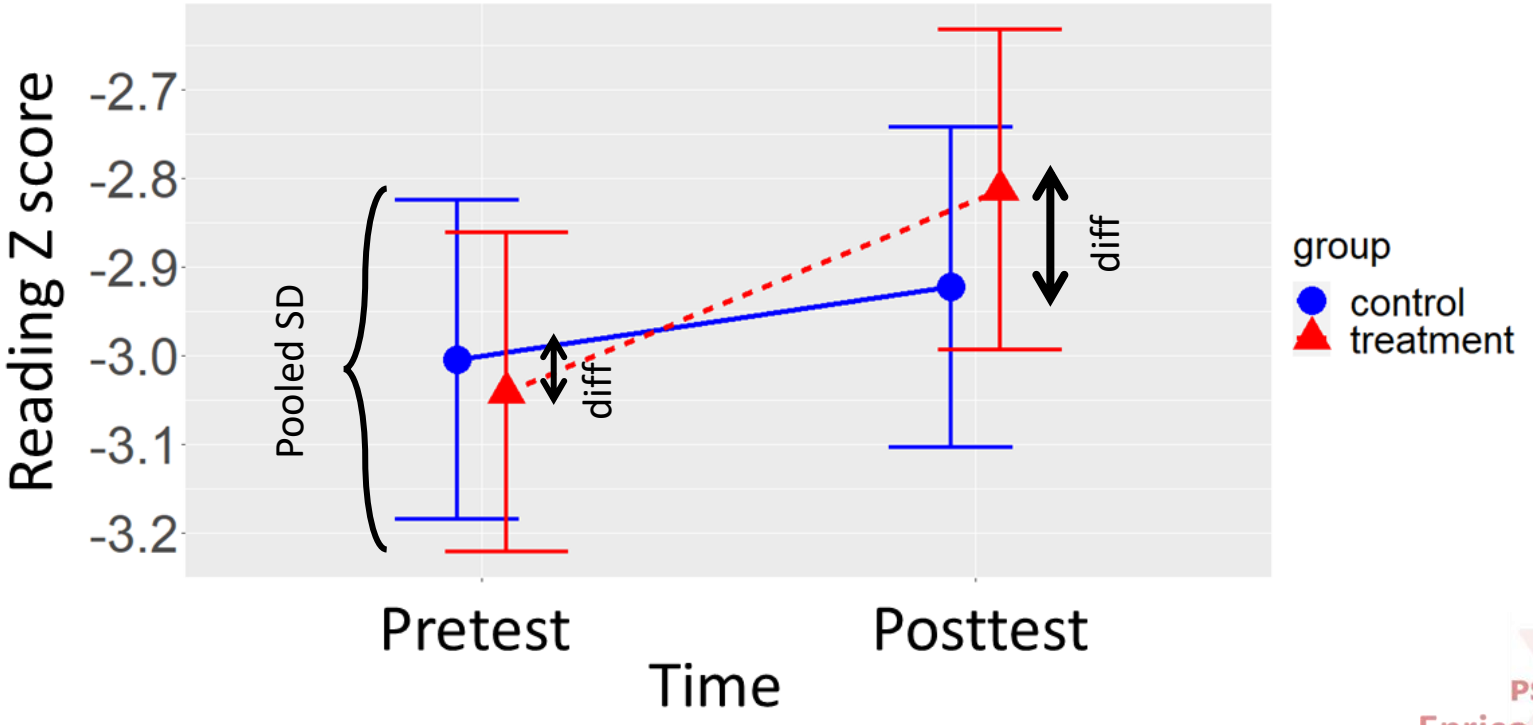
- ANOVA / mixed-effect LM with interaction

lmer(score ~ **group * time** + (1|id))

Randomized controlled trial (RCT)

Effect size *(Morris, 2008)*

- [recommended] Standardized difference pre-post in the treated group adjusted by the same in the control group (using only pre-test data for pooled SD)



Meta-analysis of 22 RCT by Galuschka et al. (2014)

- Global mean effect, **d = .30**, 95% BCI (.18, .42), **τ = .08** (recalculated by us with "brms" in R)
- For the most used and only "significant" approach (phonics training), **d = .32**, 95% CI (.18, .47), but after adjusting for ***publication bias*** → **d = .19**, 95% CI (.04, .36)
- Mean **N per group = 28**, median = **20**

OPEN ACCESS Freely available online

PLOS ONE

Effectiveness of Treatment Approaches for Children and Adolescents with Reading Disabilities: A Meta-Analysis of Randomized Controlled Trials

Katharina Galuschka[1], Elena Ise[2], Kathrin Krick[1], Gerd Schulte-Körne[1]*

1 Department of Child and Adolescent Psychiatry and Psychotherapy, University of Munich, Munich, Germany, **2** Department of Child and Adolescent Psychiatry and Psychotherapy, University of Cologne, Cologne, Germany

Abstract

Children and adolescents with reading disabilities experience a significant impairment in the acquisition of reading and spelling skills. Given the emotional and academic consequences for children with persistent reading disorders, evidence-based interventions are critically needed. The present meta-analysis extracts the results of all available randomized controlled trials. The aims were to determine the effectiveness of different treatment approaches and the impact of various factors on the efficacy of interventions. The literature search for published randomized-controlled trials comprised an electronic search in the databases ERIC, PsycINFO, PubMed, and Cochrane, and an examination of bibliographical

Meta-analysis of 22 RCT by Galuschka et al. (2014)

Insufficient power

The median study (N per group = 20) …

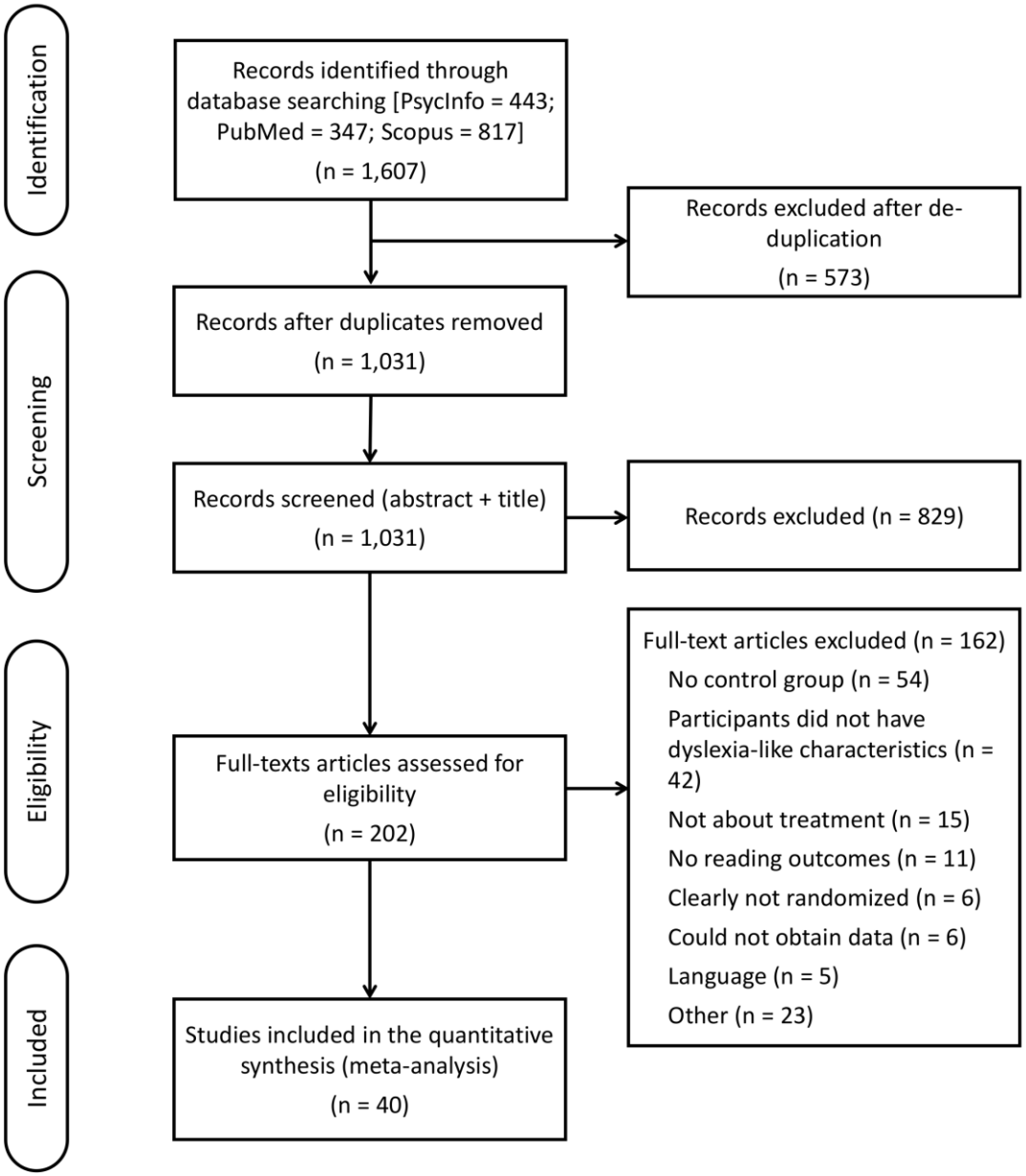
… assuming a true effect of d = .30 …

… a pretest-posttest correlation of r = .80 (strong but still plausible) … has:

Power = 34%

Exaggeration ratio = 1.67

Study 1 – Update previous meta-analysis 2013-2020



Search from APA PsycInfo, Scopus, Pubmed

40 studies included !

PSICOSTAT
Enrico Toffalini

Study 1 – Update previous meta-analysis 2013-2020

- Overall effect, calculated with multi-level random-effects model via “brms”, using Galuschka et al. (2014) data for *informed priors*, **d = .38**, 95% BCI (.31, .46), **τ = .12**, 95% BCI (.02, .24)

- Mean **N per group = 20**, median = **15**

- Too many different approaches and few studies to estimate each reliably, but heterogeneity remains

Insufficient power (again!)

The median study (N per group = 15) …

… assuming a true effect of d = .38 …

… a pretest-posttest correlation of r = .80 (strong but still plausible) … has:

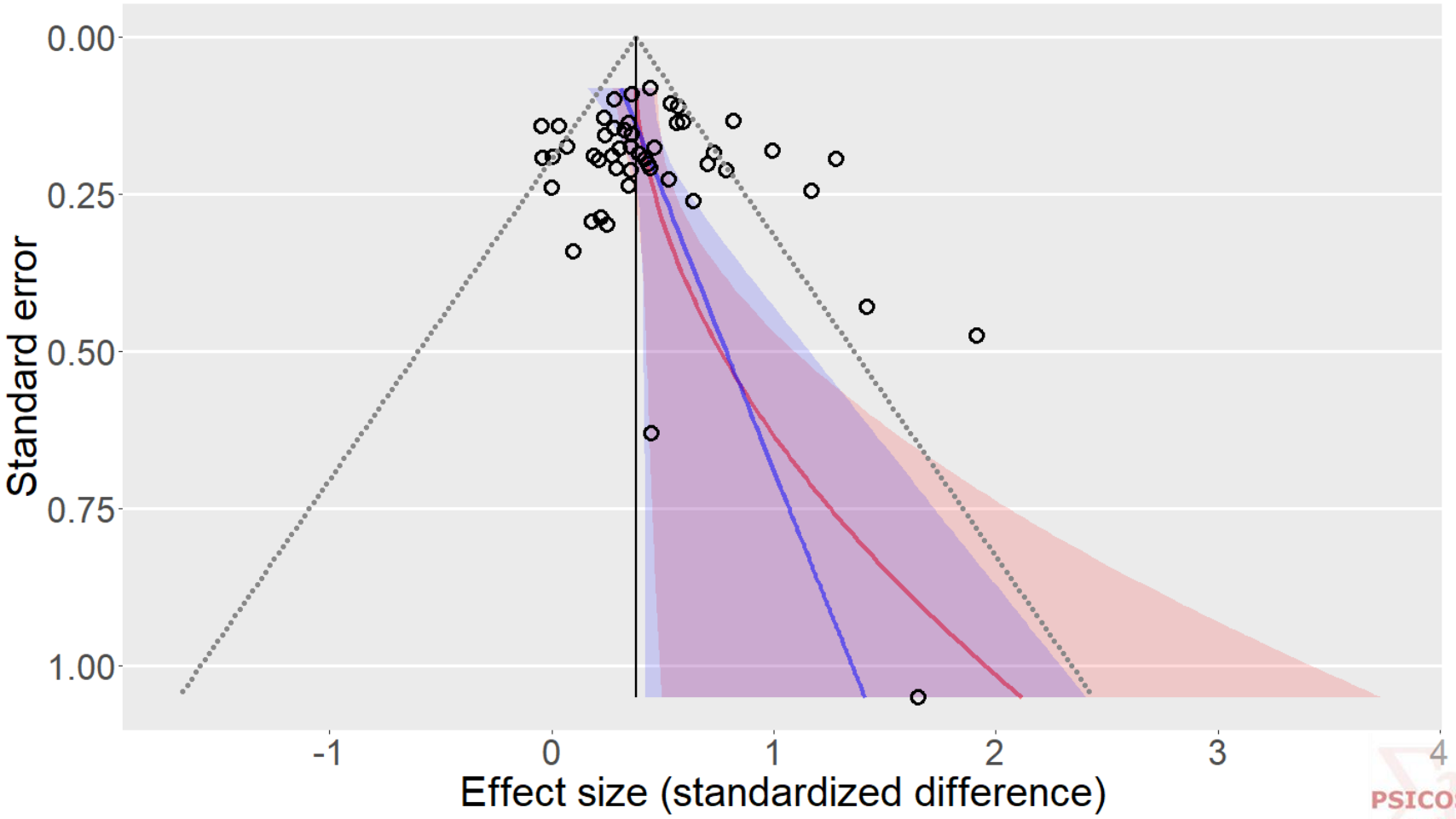
Power = 37%

Exaggeration ratio = 1.58

Study 1 – Update previous meta-analysis 2013-2020

Publication bias? Likely yes

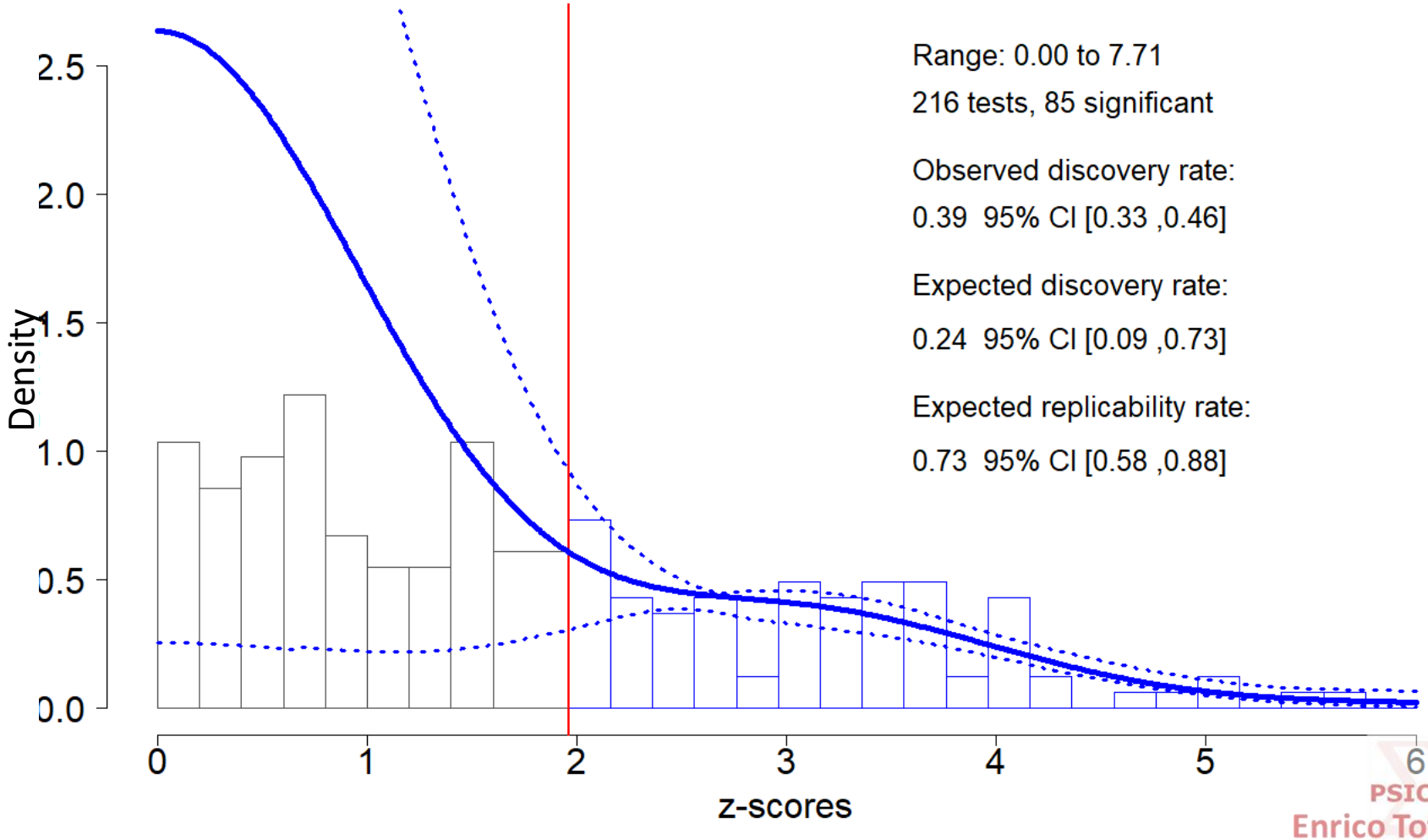
PET-PEESE model suggests asymmetry and corrects by ~10-50%



Study 1 – Update previous meta-analysis 2013-2020

Publication bias? Likely yes

Z-CURVE (on all p-values) suggests 24% expected discovery rate vs 39% observed



Insufficient power → most studies bound to fail to provide enough evidence of treatment efficacy

Why did researches insist on conducting studies **in the same way** across over more than three decades?

Most claim success

Out of all abstracts (62 [22+40] studies) :

→ 50 (81%) say reading improved!

→ 9 (15%) say inconsistent results or improvements in other areas

→ only 2 admits no efficacy

→ 1 made no claim in the abstract (focus on methods)

Study	Claim in the abstract	Text from abstract
Bhattacharya 2004	Effective!	Posttests revealed that graphosyllabic instruction helped students to decode novel words, remember how to read words with practice, and remember the spellings of words when compared to controls
Bull 2007	Not effective on reading, but on other areas…	There were no statistically significant improvements in cognitive or Literacy test performance associated with the treatment. However, there were statistically significant improvements in academic self-esteem, and reading self-esteem, for the treatment group
del Ros O Gonzalez 2002	Effective!	The results indicated that both experimental groups improved in phonemic awareness compared to the control group but that only the SP/LPA group scored higher than the control group in reading
Heikkila 2013	Effective!	Our results, based on a sample of 150 poor readers of Finnish, showed clear gains in reading speed regarding all trained syllables, but a transfer effect to the word level was evident only in the case of long infrequent syllables
Ianni 1985	Effective!	Children treated with piracetam showed improvements in reading speed
Jimenez 2007	Effective!	The results indicate that experimental groups who participated in the phoneme and syllable conditions improved their word recognition in comparison with the control group
Kirk 2009	Effective!	Participants in the experimental group made significantly greater gains in reading and spelling accuracy than those in the control group on both experimental and standardized measures of reading and spelling
Lovett 1989	Effective!	Effects specific to each experimental treatment were identified, as well as some generalized treatment advantages shared by both experimental groups at post-test. These results indicate that some of the deficits associated with developmental dyslexia are amenable to treatment.
Lovett 1990	Effective!	The word-training groups made significant gains in word recognition accuracy and speed and in spelling. Significant transfer was observed on uninstructed spelling content but not on uninstructed reading vocabulary. In general, the word-training programs were equally effective for instructive content, but the whole-word group was superior on some transfer measures at posttest
Lovett 1996	Effective!	Both the ‘‘knowledge-base’’ and the ‘‘strategy’’ training approaches were associated with significant improvement in disabled readers’ comprehension skills, although training effects did not generalize across all aspects of reading comprehension performance. Strategy-trained readers applied the trained strategies with equal success on instructed and uninstructed text materials, providing strong evidence of transfer of learning
Lovett 1997	Effective!	Both training approaches were associated with significant improvement in word identification and word attack skills and sizeable transfer-of-training effects.
Lovett 2000	Effective!	There were generalized treatment effects on standardized measures of word identification, passage comprehension, and nonword reading
McPhilipps 2000	Effective! (but on underlying "motor control"; reading was not tested)	The experimental group showed a significant decrease in the level of persistent reflex over the course of the study
Mitchell 2008	Inconsistent…	At posttest the experimental group reported statistically significantly fewer visual discomfort symptoms. The remainder of the results were, however, inconsistent
Murphy 2011	Effective!	The group receiving nonverbal auditory training demonstrated significant improvements (mainly for the group from 7 to 10 years old), not only in the nonverbal auditory skills trained (p ! 0.001), but also in phonological awareness syllable tasks (synthesis, segmentation, manipulation and syllable transposition) in experiment 1 (p = 0.003), and phonemic tasks (p < 0.001) and text reading (p < 0.001) in experiment 2.
OShaughnessy 2000	Effective!	Results indicate that children in both reading programs achieved significant gains in beginning reading skills, learning the specific skills taught in their respective programs, and applying what they had learned to uninstructed material on several transfer-of-learning measures, in comparison to children in the control group.

Robinson 1999	NOT effective	There was a significant improvement for all groups in the accuracy of miscues over the period, although experimental groups over-all did not improve ar a significantly ddIfferent rate than the control group.
Ryder 2008	Effective!	Posttests results showed that the intervention group significantly outperformed the control group on measures of phonemic awareness, pseudoword decoding, context free word recognition, and reading comprehension. Two-year follow-up data indicated that the positive effects of the intervention program were not only maintained but had generalized to word recognition accuracy in connected text.
Sanchez 1991	Effective!	After the training program, children who had received the first two training programs reached a level of performance similar to that of normal readers in different tasks of segmentation of phonemes
Tormanen 2009	Effective! (but only in "some" scores)	There was an improvement in auditory-visual matching during the training period. There were also improvements in some reading test scores, especially in reading nonsense words and in reading speed
Tressoldi 2000	Effective!	The treatment deriving from dual-route models produced significant improvements in the homophone recognition, compared to all other treatments. The treatment deriving from single-route models produced significant improvements, compared to all other treatments, in speed of word reading. Furthermore, these two treatments produced significant improvement with respect to all other treatments but one, in speed of nonword reading.
Wilsher 1987	Effective!	Piracetam-treated children showed significant improvements in reading ability (Gray Oral Reading Test) and reading comprehension (Gilmore Oral Reading Test). Treatment effects were evident after 12 weeks and were sustained for the total period (36 weeks).

examples from Galuschka et al (2014)

How is it possible?
You know how…

CONVENIENT MULTIPLE TESTING

In meta-analyses, both Galuschka et al. (2014) and us appropriately combined outcomes within the same study and group comparison

However, treatment studies claimed success based on $p < .05$ for even just one single outcome, in just one comparison (sometimes playing with covariates)

→ on average, each study tested 4.5 outcomes (median: 3-4 outcomes)

only ~ 25% studies applied any p-value correction for multiple tests, but almost never for multiple outcomes

Study 2 – Suggestions on how to improve power

Reliability is a key for precision (and power)

Variations of a reading outcome reflect:

- *Response to treatment (in treated individuals)*

} EFFECT / SIGNAL

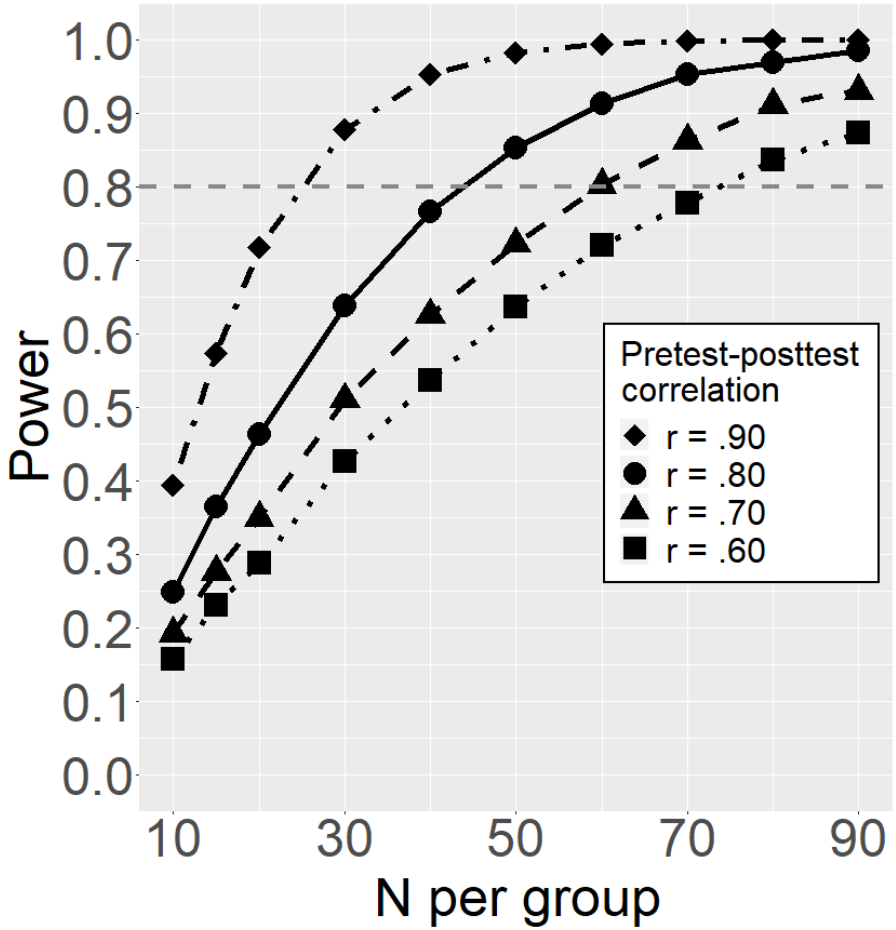
- ***Measurement error*** ← good **reliability** minimizes this
- *(Different) developmental trajectories over the time of the study* ← hope this is small

} ERROR VARIANCE / NOISE

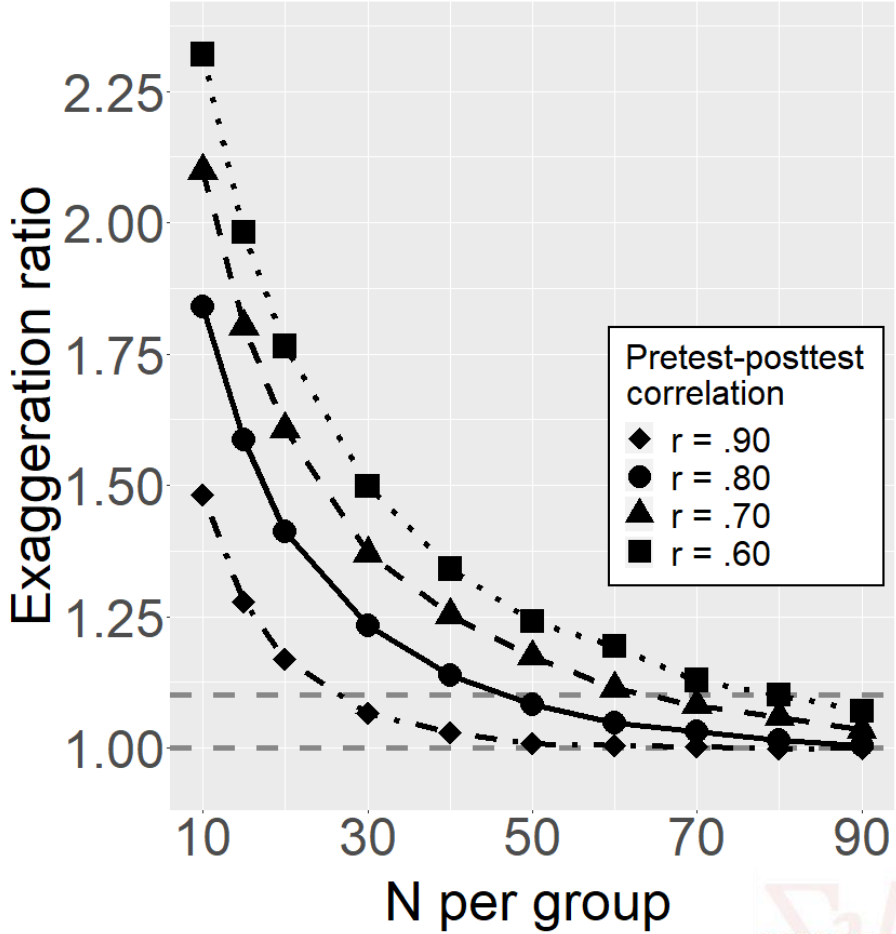
Study 2 – Suggestions on how to improve power

Design analysis via simulation for **d = .38** (±.20 between-subj)

(A) Power



(B) Exaggeration ratio



Study 2 – Suggestions on how to improve power

Design analysis via simulation for **d = .38** (±.20 between-subj)

With a plausible pretest-posttest r = .70-.80, the adequate N per group is 50-60 (total sample size at least 100-120), for power 80%, type M error ~10%

This requires **massive** effort!

Study 2 – Suggestions on how to improve power

What is the real stability of my measure?

- Stability of reading measures (e.g., word lists) often very high, r ~ .85-.90 from standardized batteries covered in our meta-analysis

- This refers to the exact same task administered twice, different versions may show slightly lower correlations

- Reliability may be lower in dyslexia population, due to the shrinkage of range of scores and/or clinical reasons; r ~ .70-.81 (Cirino et al., 2002; Wang, 2017; Wang et al., 2019)

Study 2 – Suggestions on how to improve power

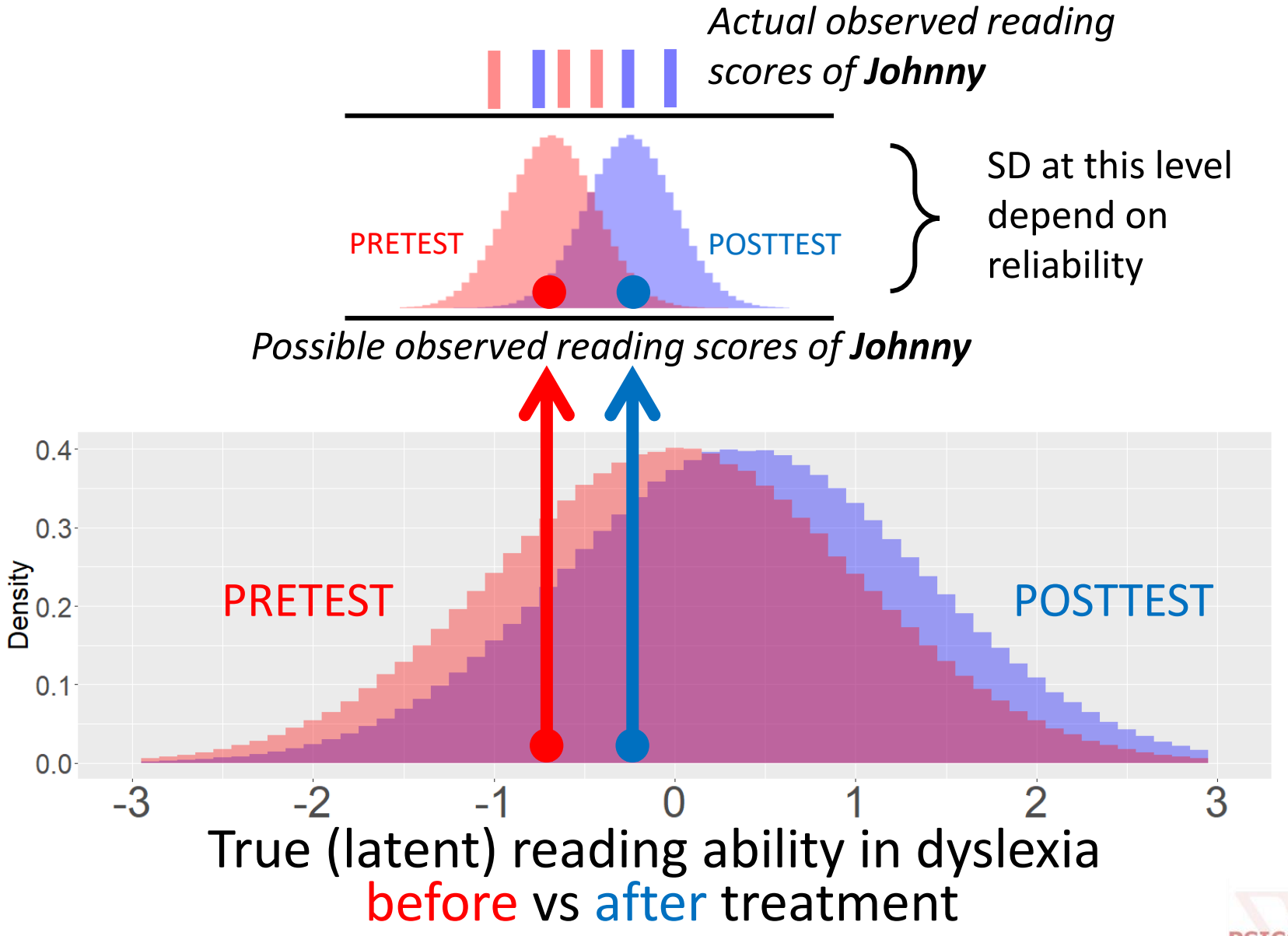
Independently from the reliability of your measure, you can improve precision by

collecting several measurements

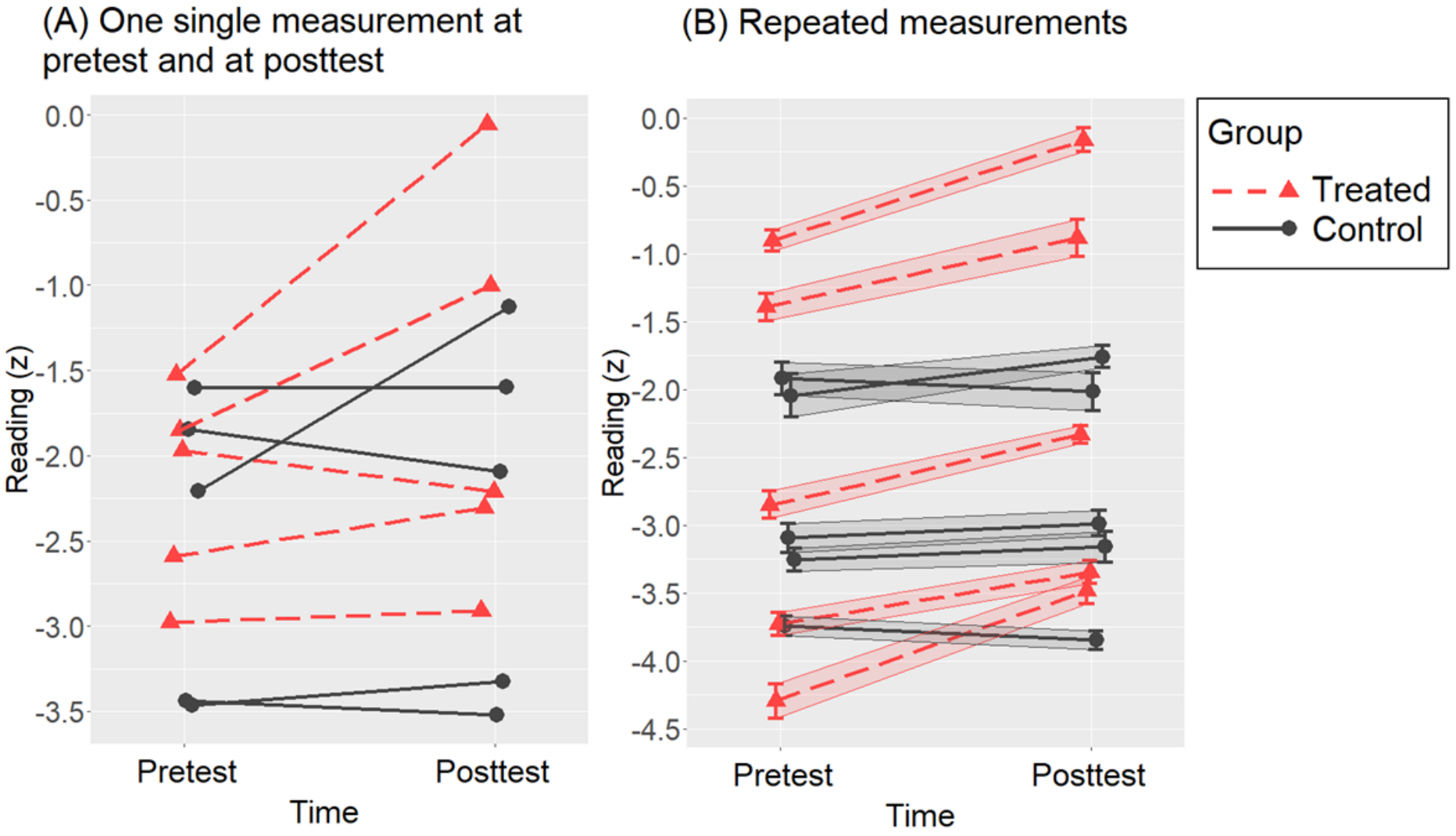
As for reading, it is easy to create multiple parallel versions of tasks (e.g., word lists)

In a dyslexia treatment study, collecting 3 measurements per time point (instead of 1), requires modest additional effort and brings huge advantage

Repeated measurements



The advantage of measuring with precision



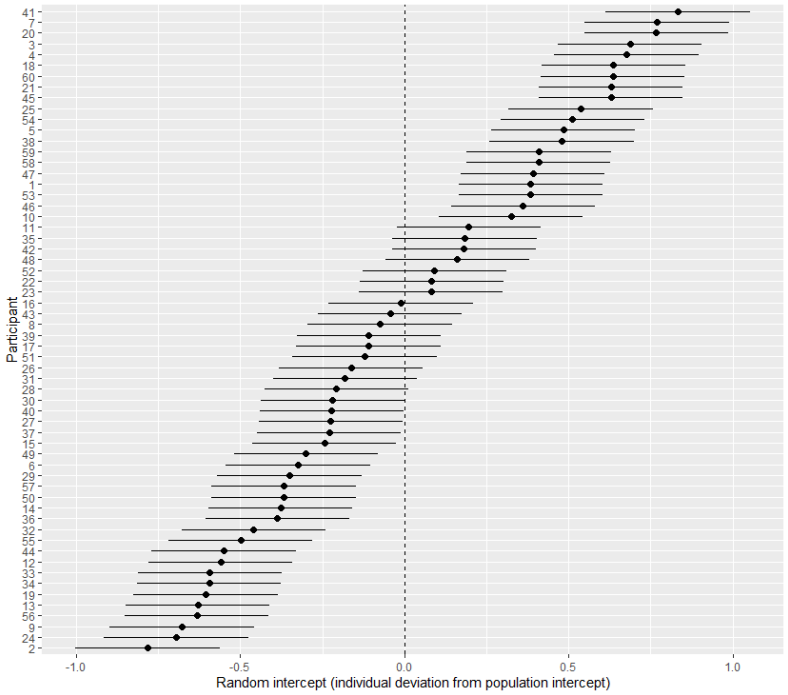
Modelling

Not actually a complicated mixed-effects model

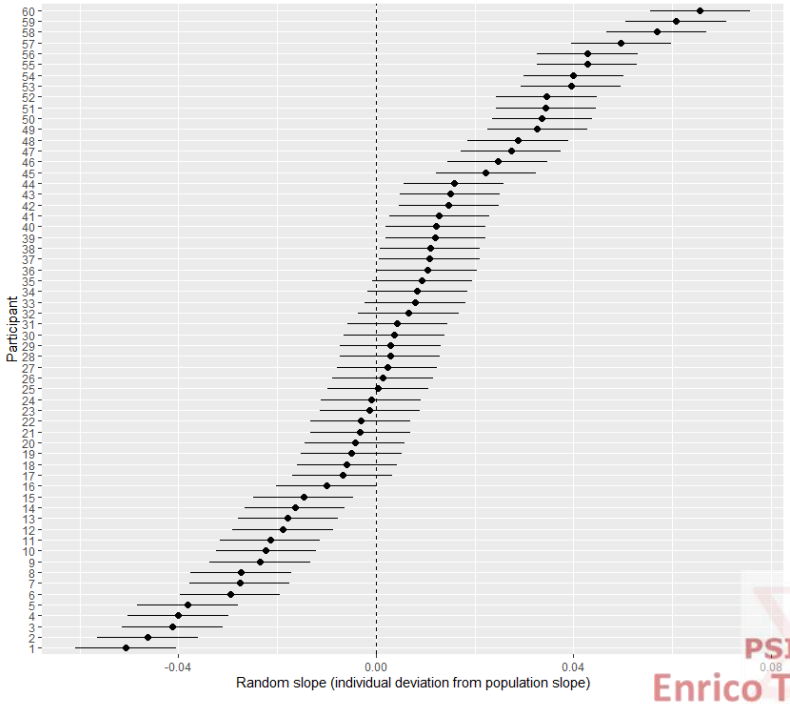
fit <- lmer (score ~ time * group + (time|ID), data = d)

random slopes may be difficult to estimate with only 3 observations at pretest and 3 at posttest!

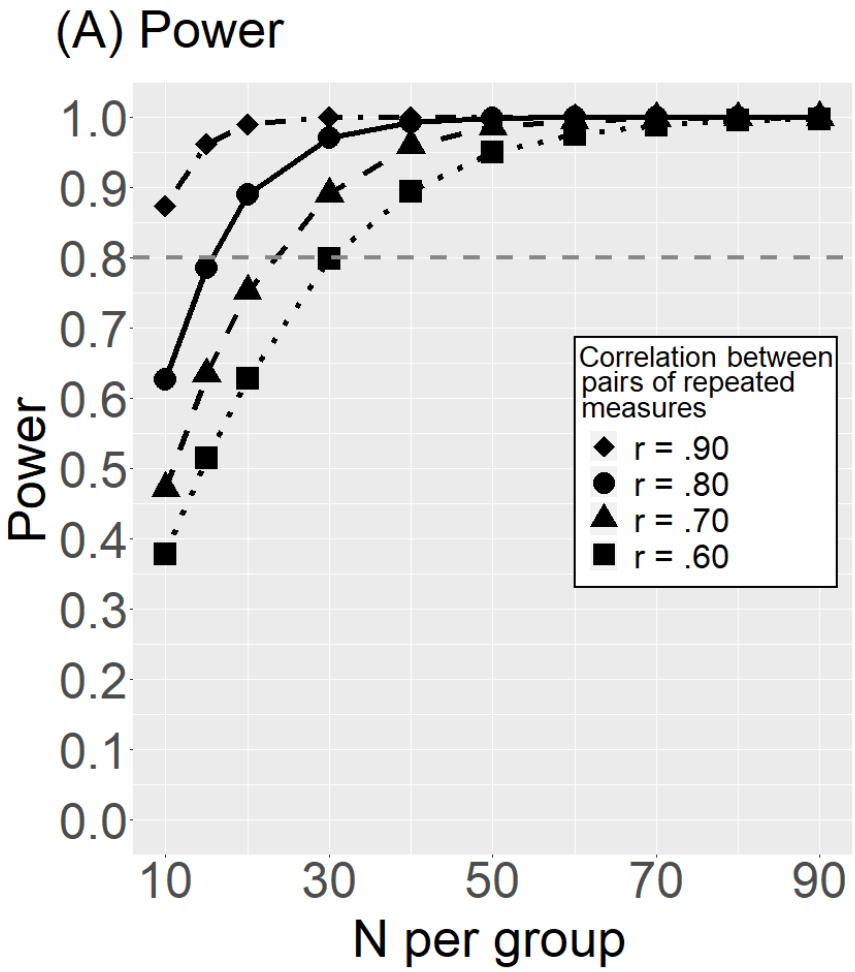
***Random intercepts*: Estimated individual differences in baseline performance**

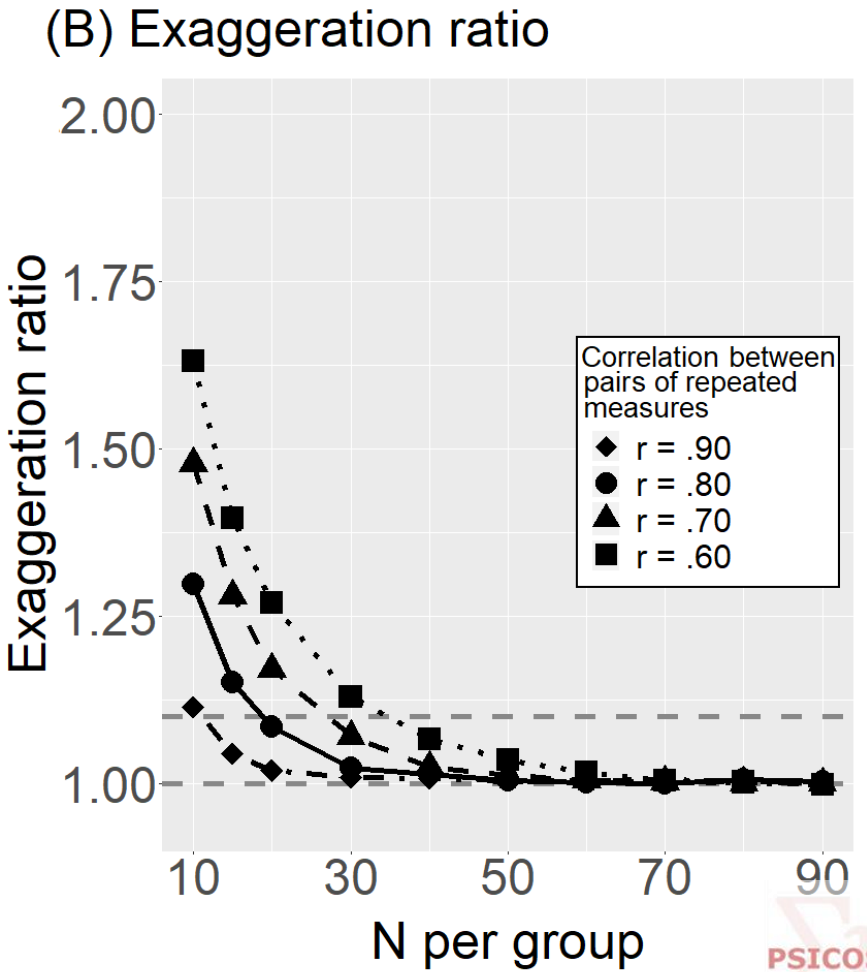


***Random slope*: Estimated individual differences in response to treatment**



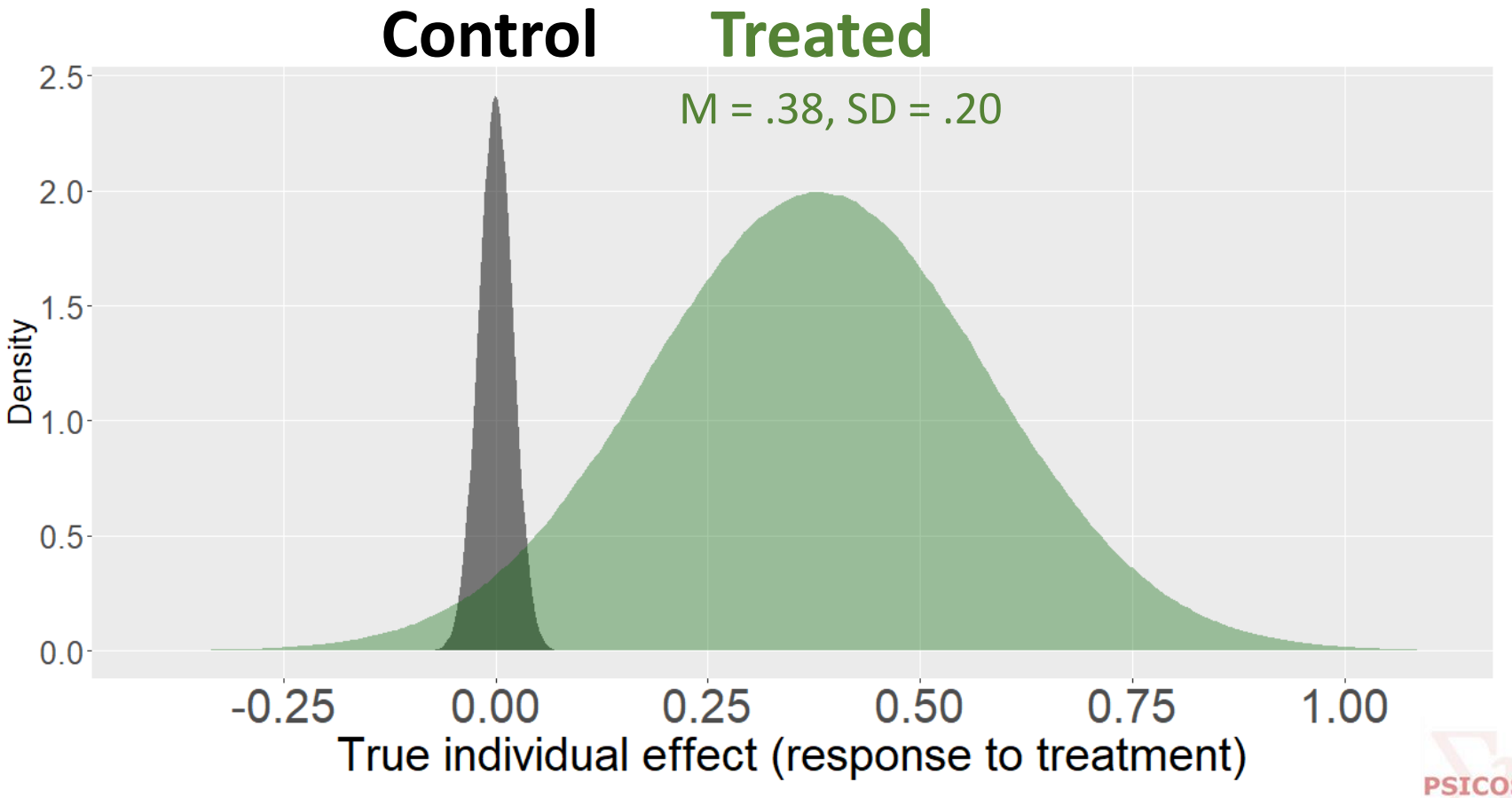
With the same assumptions as above, but 3 repeated measurements per time point, the required N per group drops from 50-60 to just about 25





Variability in response to treatment? Why not!

It is part of what can be of interest, and it can be investigated, so it can be considered in design analysis

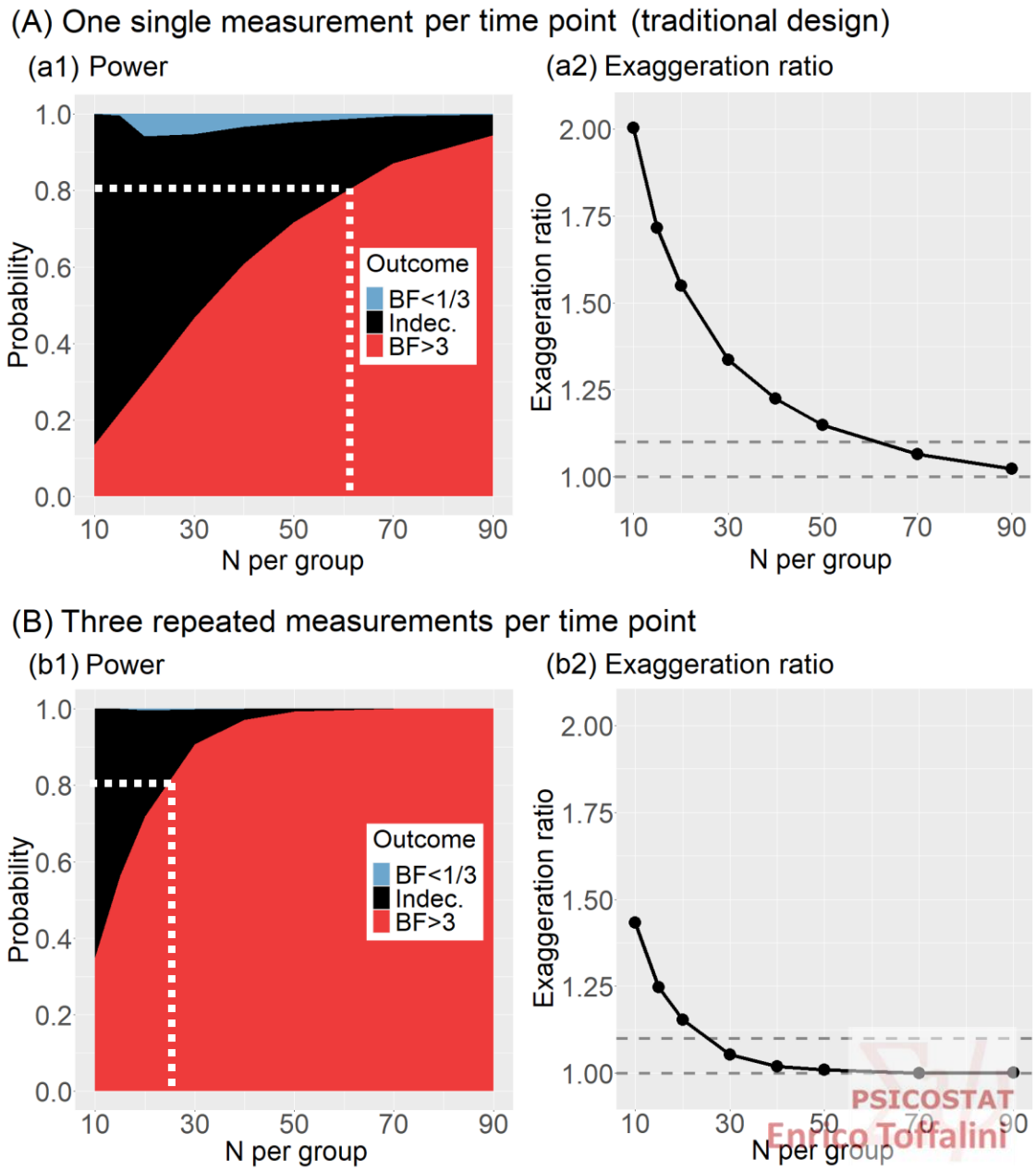


Bayes Factor as inferential criterion?

d = .38, r = .80

again, very good choice to use repeated measurements!

But not better than the frequentist method, especially if you mindlessly use default settings and interpretations



Conclusions

- Assessing treatment efficacy is difficult: Effect size is modest, sample size an issue, publication bias likely

- Strategic use of repeated measurements can be the key for precision, power, and individual differences
  - Timing of measurement matters! Try to capture maximum variability
  - Beware! Power cannot be increased indefinitely with repeated measurements: with too few participants you cannot generalize as you remain uncertain about real variability in response to treatment
  - But hey, anything can be considered *a priori* with a design analysis!

- Default Bayes Factor won't solve your problems

Thank you for your attention