August 2026

HOW TO INTERPRET EFFECT SIZES AND P-VALUES

What Constitutes Valid Research

by Dr Sharon McMurray MBE

Effect sizes

> *An effect size is a statistical number that shows the size or strength of a difference between groups or a relationship between variables. While a p-value tells you if a result is due to chance, an effect size tells you how meaningful or large that result is in the real world. Effect sizes matter because they show the importance in practice and allow researchers to compare results across different studies. Large sample sizes can make tiny, unimportant differences look statistically significant; effect sizes show if the actual difference matters.*

0.2	Small effect
0.5	Medium effect
0.8	Large effect
1.0+	Very large effect

P-values

> *P-values provide a measure of the extent to which the result is due to chance.*

P-VALUE	CHANCE THE RESULT IS RANDOM	SIGNIFICANCE LEVEL	STRENGTH OF EVIDENCE
$p < 0.05$	Less than 5%	Statistically significant	Moderate evidence
$p < 0.01$	Less than 1%	Highly significant	Strong evidence
$p < 0.001$	Less than 0.1%	Very highly significant	Very strong evidence
$p < 0.0001$	Less than 0.01%	Extremely significant	Extremely strong evidence

Sample size isn't everything

A larger sample size does not make evaluative research more reliable or methodologically stronger than a smaller well-designed study. Sample size increases statistical power, but it does not compensate for poor methodology or bias.

The larger the research sample does not mean stronger evidence. Maintaining consistent procedures across a very large sample can introduce additional methodological challenges — for example, ensuring consistent delivery, assessment, participant selection, data collection and researcher training. Those who claim that larger research studies are better are expressing a very simplistic view. The strength of research evidence should be judged primarily by the methodological rigour of the study, rather

than sample size alone. A large sample cannot compensate for fundamental weaknesses in research design or systematic bias.

What makes research methodologically sound?

Rigorous research studies that evaluate educational interventions should meet at least six of the eight indicators listed below to be considered methodologically sound or rigorous:

- The use of a comparison control group(s) and experimental/treatment group(s).
- The random assignment of participants to experimental/treatment or control groups.
- The establishment of baseline equivalence between groups.
- The inclusion of sufficient information regarding the participant sample for study replication.
- The use of standardised scores (age-equivalent scores such as reading and spelling ages should not be used as they are unreliable).
- Enough information to calculate effect sizes.
- The report of the fidelity of treatment.
- The use of blind evaluators.

> **SIX IS THE THRESHOLD**
>
> Six or more is the threshold for rigour. Having a comparison/control group, using standardised scores and reporting effect sizes are considered essential for rigour. None of the 8 indicators should be omitted if it is possible within the research design. Blind evaluators are particularly important to ensure there is no bias. A strong methodological rationale must be given for any indicator that is not included, and the threshold of 6 must be met.

The CSP Spelling and Language Programme: a case study in rigour

Taking my PhD research on the CSP Spelling and Language Programme: The Complete Spelling Programme as an example, the core criteria to ensure rigour in this PhD research was:

- Schools were randomly assigned to the control and experimental groups based on criteria to establish baseline equivalence (RCT). Following this a 3-year quasi-experimental longitudinal research design was implemented.
- Information on participants was provided for study replication.
- Standardised scores were used (reading ages and spelling ages, i.e., age-equivalent scores, would not meet the requirement for rigour).
- Fidelity of implementation was very closely monitored.
- Tests were administered by research assistants, and two research assistants were in every test session — one to administer the tests and the second to ensure that the tests were administered correctly with strict adherence to test procedures and wording.
- The research assistants marking and cross-marking the standardised tests were not the research assistants who administered the tests, and did not know whether the test papers they were marking were from control or experimental schools (a requirement for blind evaluation).

- The assessment of spelling in independent writing samples, and the evaluation of the quality of independent writing samples, was completed by experienced teachers, and inter-rater reliability was established. These teachers did not know whether the samples of writing were from the control or experimental schools (a requirement for blind evaluation).

To ensure the validity of the data, and protect the lead author of the intervention from any future criticism, this was a rigorously controlled study. I did not administer the tests, collect any of the data, mark any of the standardised tests or evaluate the samples of writing. The data was entered into SPSS by staff in the School of Education at Queen's University Belfast to safeguard its integrity, and they ran the analyses.

> *"Remarkable" — Professor Greg Brooks (2013), describing the results of the CSP Spelling and Language Programme (p<0.0001, effect size 1.19), reported in "What Works for Literacy Difficulties".*
> THESCHOOLPSYCHOLOGYSERVICE.COM/WHAT-WORKS/CSP, ACCESSED 13/08/2026

REFERENCES

Brooks, G. (2013). What works for children and young people with literacy difficulties? The School Psychology Service. theschoolpsychologyservice.com/what-works/csp