EXAMINING THE EVIDENCE BASE

SSP Research Evidence: Johnston & Watson (2005)

by Dr Sharon McMurray MBE

The lack of robust research to support the exclusive use of SSP over any other form of systematic phonics is evidenced by the use of the Johnston and Watson (2005) study by those promoting SSP. After over 20 years of using SSP in England no valid research has been identified. However, the flaws in the Johnston and Watson (2005) study are extensive and make the findings unreliable. This paper considers the three main flaws.

1. There was no control group beyond the first 16 weeks

After the initial 16-week experiment, both analytic groups were switched to synthetic phonics by the end of Primary 1. That means the 7-year follow-up did not compare whether synthetic phonics produced better long-term outcomes than continuing with analytic phonics, and the claims made are not valid.

2. The comparison is timed unfairly

Analytic phonics is described by the authors themselves as a slower, cumulative programme that only introduces letter-sound correspondences in all positions late in Primary 1 and completes by Primary 3. Testing everyone at 16 weeks compares the synthetic group after it had been taught most grapheme-phoneme correspondences against analytic groups that had deliberately not yet reached that point. The curricular content at the comparison point is not equated.

3. Standardised scores were not used

The single biggest analytic problem running through the Johnston & Watson report is that almost every headline reports "months ahead", e.g., "3 years 6 months ahead", and these have been reported in DE (2025), page 42. Months ahead are calculated by taking arithmetic means of age-equivalent scores (reading ages, spelling ages, comprehension ages) and by subtracting chronological ages from them. Reading ages and spelling ages are not interval or ratio scales of measurement and should not be added, subtracted or averaged, and are considered to be so unreliable that they should not be used in any decision making. In addition to the unreliability of reading ages and spelling ages, reporting using "months ahead of chronological age" as the headline outcome exaggerates differences that are in themselves unreliable.

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.

IN CONCLUSION

This Johnston and Watson research is not methodologically sound and should not be used as support for systematic synthetic phonics.

HOW TO INTERPRET EFFECT SIZES AND P-VALUES

What Constitutes Valid Research

by Dr Sharon McMurray MBE

Effect sizes

Effect sizes show the magnitude of the difference between groups. The effect size reflects the strength of the intervention's impact. It shows how meaningful or substantial the effect is in practical terms.

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 < 0.05	Less than a 5% chance the result is random. **Statistically significant** — moderate evidence.
p < 0.01	Less than a 1% chance the result is random. **Highly significant** — strong evidence.
p < 0.001	Less than a 0.1% chance the result is random. **Very highly significant** — very strong evidence.
p < 0.0001	Less than a 0.01% chance the result is random. **Extremely significant** — representing 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.

- ## 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

Johnston, R., & Watson, J. (2005). *The effects of synthetic phonics teaching on reading and spelling attainment: A seven year longitudinal study.* Scottish Executive Education Department.

Department of Education (DE). (2025). *Strong foundations: A literacy framework for primary schools.*

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