



What counts as ‘evidence’ in literacy education?

Beverly Derewianka¹ · Helen Harper² · Bronwyn Parkin³ · Claire Acevedo^{4,5} · David Rose⁶ · Brian Dare⁷ · Maria Estela Brisk⁸ · Pauline Jones¹

Received: 26 August 2024 / Accepted: 11 September 2024 / Published online: 26 September 2024
© The Author(s) 2024

Abstract

A recent issue of the Australian Journal of Language and Literacy included an article reporting on a systematic narrative review of the research literature that indicated that there was insufficient evidence to conclude whether genre theory and systemic functional linguistics either ‘worked’ or ‘did not work’. The criteria used to evaluate these studies excluded any study that did not conform to the ‘gold standard’ associated with experimental research such as randomised controlled trials. In response to this provocative finding, a group of SFL researchers decided to examine just what counts as evidence of quality literacy research these days. In this paper, we question the overreliance on experimental research at the expense of other methods. We illustrate this with a sample of notable studies that do not meet experimental criteria, but which nevertheless have made a significant contribution to school literacy outcomes in Australia and elsewhere.

Keywords Systemic functional linguistics · Genre theory · Literacy education · Evidence-based research

✉ Pauline Jones
paulinej@uow.edu.au

¹ University of Wollongong, Wollongong, NSW, Australia

² University of New England, Armidale, NSW, Australia

³ University of South Australia, Adelaide, SA, Australia

⁴ London, UK

⁵ ForMuLE Complutense University of Madrid, Madrid, Spain

⁶ University of Sydney, Sydney, NSW, Australia

⁷ Lexis Education, Adelaide, SA, Australia

⁸ Boston College, Boston, USA

1 Introduction

As a group of educational linguists,¹ we were intrigued by the publication in a previous issue of *The Australian Journal of Language and Literacy* (AJLL) of an article reporting on a systematic review of the effectiveness of literacy practices that draw on genre theory (GT) and/or systemic functional linguistics (SFL) (Green et al., 2024). The article identified some 7846 peer-reviewed journal articles, books, chapters, and research reports dealing with reading and writing outcomes in grades K-10 from Australia, Canada, UK, New Zealand and the USA (thereby excluding research studies from areas where GT/SFL is increasingly strong, such as Scandinavia, China, Indonesia and South America). Surprisingly, it also excluded any that dealt with English as a second/additional language, Indigenous literacies, and multimodal texts—areas of particular interest in GT/SFL research.

Beyond these exclusions, Green and colleagues eliminated from the review process any studies that did not meet the following criteria:

- Peer-reviewed intervention studies
- Quantitatively measured outcomes
- Experimental/quasi-experimental methodology
- Contained a control/comparison group

Of the 7846 studies, only 6 were deemed to have met the criteria to a certain extent.

For those 6 studies that qualified for review, the criteria for evaluating study quality included evidence of the following:

- Random sampling and assignment
- Comparability of sample and control groups
- Adequate sample sizes
- Composition of participants (e.g. any language disorders, behavioural issues)
- Reporting quality (e.g. implementation frequency, number of participants)
- Detailed description of comparison instruction/intervention
- Fidelity of implementation of intervention
- Any issues with attrition
- Quantitatively measured outcomes
- Reliability and validity of measures
- Statistical data on concurrent validity
- Source and quality of measurement instrument (e.g. external standardised test)
- Rigorous calculation of measurement of outcomes (e.g. effect sizes, growth rates)
- The use of multiple outcome measures used to balance between measures aligned with the intervention and generalised performance

Our intention here is not to respond in any detail to the article in question, but to address broader issues it raises with respect to what counts as evidence of quality research in literacy education these days. To this end, we will briefly introduce evidence-based instruction, summarise some of the issues with experimental and

¹ This response was prompted by a lengthy email exchange involving some 30 international educational linguists.

quasi-experimental research as the main source of evidence, consider its impact on policy and decision-making, propose some alternatives, and exemplify these with a sampling of studies from the GT/SFL literature rejected by Green and colleagues.

2 Evidence-based instruction

The current preoccupation with ‘evidence-based’ teaching practices had its roots in the late twentieth century as a backlash against constructivist approaches of the time that were characterised as student-led discovery learning (Thomas, 2016). In an attempt to make educational research more ‘scientific’, the randomised controlled trial (RCT), grounded in clinical research, was elevated as the gold standard of research design for assessing the impact of interventions. The questionable implication is that because research emerging from any other design is deemed to lack the rigor of RCT-based research, it therefore cannot reliably tell us ‘what works’ (Mosteller & Boruch, 2002; Parra & Edwards, 2024).

In their AJLL article, Green and colleagues reiterate the conviction that the only research to be considered as evidence of ‘what works’ is that which conforms to ‘the gold standard’, preferably in the form of RCTs. The Grattan Institute, for example, has called for Australia to lift the standards for scientific evidence in education and produce more randomised controlled trials and quasi-experimental studies (Hunter, 2022). Further, the Australian Educational Research Organisation (AERO) urges Australia to build on the evidence base of what works, providing a clear rationale for why strategy *x* produces result *y* that is measurable and ‘done in the most robust way possible (including use of randomised controlled trials where appropriate)’ (AERO, 2023, p. 6).

Typical features of RCTs in education include measurable outcomes from an intervention, experimental design, narrow-angle lens, longitudinal studies, large sample sizes (the ‘law of large numbers’), double-blind studies, randomised comparison groups from more than one institution, statistical data analyses, intervention effect sizes, and generalisable findings of predictable human behaviour (Hubbard et al., 2019; Parra & Edwards, 2024; Sedgwick, 2015).

However, as McKnight and Morgan (2022) caution, ‘it is naive to transpose the gold standard for specific kinds of research in medicine onto an entirely different discipline, such as teaching’. More radically, philosopher of science and knowledge Michael Scriven proposes that ‘The RCT design... has essentially zero practical application to the field of human affairs’ (Scriven, 2008, p. 12).

3 Issues with ‘the gold standard’ in education

Over the past few decades, concerns have been raised about the dominance of experimental studies in education as the arbiters of what counts as evidence. Parra and Edwards (2024), for example, note ‘the exponential growth and high-profile endorsement of RCTs to inform evidence-based education in the last two decades’, citing Barrenechea et al. (2023), Connolly et al. (2018), and Simpson (2019). Here, we will review some of the issues concerning the ascendancy of experimental research—and RCTs in particular—in education.

3.1 Complexity/context

Perhaps the most common critique is that education is a highly complex system with multiple interacting components, defying the RCT's need for tightly controlled variables. RCTs work best when the research 'confers a high probability on the causal conclusion' (Cartwright, 2009, p. 60), yet the complexity of education reduces the ability to attribute causality to activity.

It is clear that RCTs might cope, for example, with constrained literacy skills such as phonics and spelling (Paris, 2005) or with isolated aspects of reading or writing or with some tightly defined sub-sample of learners. The argument here is that they are not designed to evaluate such instructional resources as the teaching and learning cycle commonly associated with genre pedagogy—which is intended to be flexible and context-dependent, integrating oral interaction, reading, writing and knowledge about language in the context of particular curriculum tasks and genres, across a range of, in some cases, incommensurate disciplines and with particular cohorts of students.

In summarising these concerns, Morrison and van der Werf (2012) pointed to the 'mess and mire' of education contexts and the fundamental 'person-ness' of the teaching situation, declaring that 'RCTs may be unachievable or undesirable, given that the social world is not the antiseptic, controllable world of the laboratory, but is complex, irreducible to simple components, multi-causally directional, dynamic, fluid, evolving and everchanging' (ibid p. 1).

3.2 Generalisability

Put simply, such context dependency, however empirically warranted, makes it harder to generalise from one situation to another. Rather than determining what works in one classroom setting as evidence for what works in all, RCTs may simply show surface level patterns (Tikly, 2015) of what worked for a particular sample, at a particular time and in a particular place, without being able to explain how and why, but nonetheless wanting to assert that the patterns will work equally well in other contexts under different circumstances when implemented by different people (Ginsburg & Smith, 2016; Joyce & Cartwright, 2018; Kaplan et al., 2020; Morrison & van der Werf, 2012; Pawson & Tilley, 2004). This tendency potentially prevents users of the research from understanding how to adapt the research to account for different contexts, and classrooms in particular represent a multitude of contexts. From the perspective of a class of individuals, teachers must know not only what learning strategies might generally work but must be aware of and support those for whom the strategy is not effective. Teachers deal with unique individuals and with particularity, and unthinking application of generalised findings to individuals risks not meeting the specific needs of learners.

3.3 Fidelity

Problems regarding the degree of fidelity with which the intervention was applied during the trial is a matter broadly acknowledged by researchers involved in large-scale RCTs (Koutsouris & Norwich, 2018; Sullivan, 2011). These include such contextual factors that impact on results as follows:

- The enthusiasm of teachers and pupils
- The level of experience/expertise of teachers and their in-depth familiarity with the theory and practice underpinning the intervention
- Perceived relevance
- Students' previous learning
- The involvement of the school leaders and local literacy adviser
- The ethos of the school (as it impacts on the approach to literacy provision)
- The way the form of the intervention fits with pre-existing organisational arrangements
- Adjustments to the intervention (e.g. pace, intensity, length and frequency of time spent)
- Difficulty in 'blinding' students and teachers and possible 'contamination' as participants in the comparison groups may share information
- Adaptations to local context (e.g. responding to student needs; the provision of additional input for certain pupils)
- The effects of other, non-intervention training experiences occurring during the study intervention

3.4 Ethics

Ethical considerations surround the implementation of RCTs. In particular, stakeholders often express their concern that it might be unethical to withhold the intervention from some students (Koutsouris & Norwich, 2018), with some parents protesting if their children receive an education 'placebo'. Others, however, object that some students might be subjected to an 'untested' intervention (Carolan & Zielezinski, 2019).

3.5 Feasibility

Fundamental questions have increasingly arisen regarding the feasibility of implementing RCTs in school settings (Biesta, 2010). Because such trials are inevitably large-scale, multi-site and protracted (even across several years), they are costly and require considerable expertise. They are typically restricted to implementation by large school systems or to university consortia—though research funding for education is scarce and industry support close to non-existent (Sullivan, 2011).

It is difficult to recruit and retain schools to participate in such a demanding exercise. While appreciating the benefits of such research, school leaders are wary of the potential disruption to routines and priorities and are mindful of restrictions placed on teacher workloads, including additional time for professional learning. The challenges of implementing experimental or quasi-experimental studies are even greater in educational contexts that are geographically isolated, where there is high teacher turnover and high student mobility, such as in the most remote regions of Australia (Helmer et al., 2013).

High levels of sample attrition are a major problem, jeopardising the internal validity of the trial as replacement participants are likely to systematically differ from teachers in schools that met the requirements of random assignment (Koutsouris & Norwich, 2018; Parra & Edwards, 2024; Troyer, 2022). Henneberger et al. (2023, p. 1) calculate that researchers should plan for attrition rates as high as 27% during middle school and 54% during elementary school.

In sum, RCTs (and most experimental studies);

- face challenges pertaining to the use of tightly controlled variables in the complexity of education settings
- favour research that focuses on discrete, constrained skills that are more easily generalisable to other contexts
- require a fidelity of implementation that is hard to replicate
- raise some substantial ethical considerations
- lack general feasibility due to the fact that they are typically large-scale, multi-site, protracted, disruptive and costly

4 Policy and power

The article in question stated that ‘a goal of this study is to make recommendations in this educational policy space’, highlighting the issue for us of the role of scientific trials such as RCTs in policy decision-making. Politicians and bureaucrats require compelling, evidence-based results that can be used to make decisions about public spending on research and effective practices and to determine future policy directions and funding. In this high-stakes environment, it matters whose voices are heard. The aura of certainty and generalisability offered by the sciences (and in particular, the RCT) has obvious appeal to politicians, policymakers and publishing houses, who then recontextualise the research findings into unambiguous, evidence-based programs (as ‘the science of ...’).

Such endorsement however is not without its problems, as noted by research methodology scholars in the humanities and in education. Phillips (2007) cautions that seeing certain research findings as ‘the answer’ is magical thinking. In seeking reassuring certitude, for example, the nuances, caveats, and variations within the body of the original research are often elided in the recontextualisation process, particularly when interpreted by those lacking the required expertise (Singh, 2002). Research evidence can be used selectively to support a policy or practice that has appeal on political or ideological grounds (Phillips, 2007). Knowledge produced in the confines of laboratories and tertiary institutions might not resonate with the reality of the classroom and could result in unwarranted restrictions when mandated as policy (Biesta, 2010). With respect to current debates about the teaching of literacy, Shanahan cautions against the ‘premature translation of basic research findings into wide-scale pedagogical application’ (2020, p.235), citing the tendency among some authors to use the findings of cognitive science and neuroscience to prescribe particular approaches to reading instruction.

In general, many scholars, working from a variety of theoretical and methodological positions, have argued that RCT studies are actually a poor basis for informing decision-making (Evans & Popova, 2016; Lortie-Forgues & Inglis, 2019; Simpson, 2019; Vivalta, 2020). Policy and decisionmakers need ‘to ensure their definition of evidence means quality, valid, and trustworthy evidence— not simply quantitative evidence of any level of validity and reliability’ (Shelley et al., 2009, p. 3). The question then arises as to where we should turn for the compelling forms of research that policymakers in education can use.

5 Alternatives to experimental evidence

Taking into account the increasing concern that ‘there is no evidence that reliance on RCTs has produced better outcomes [in policy-oriented research] than alternative approaches’ (Muller, 2020, p.2), perhaps it is opportune to consider other approaches for informing

evidence-based education (Barrenechea et al., 2023; Gale, 2017; Joyce & Cartwright, 2018; Tikly, 2015; Wrigley & McCusker, 2019).

5.1 Complementing the RCT

One way of adding further validity to the findings of RCTs is to accompany them with qualitative studies—a blended, mixed-method approach that overcomes the unproductive divide between quantitative and qualitative methodologies. Each source of data and methodological approach can contribute important elements of the educational effort (Malina et al., 2010; Phillips, 2007). To overcome considerations relating to the interaction of intervention components with each other and with the context, ‘realists’ argue that trials should draw on complementary qualitative and quantitative data, examining the effects of intervention components separately and in combination (Bonell et al., 2012; McKnight & Morgan, 2022).

Carolan and Zielezinski (2019) suggest an efficacy portfolio which may include an RCT but would comprise a compilation of evidence from different sources (e.g. feasibility studies, case studies, user interviews, and correlational studies). Similarly, Koutsouris and Norwich (2018) propose a process evaluation to accompany the RCT, taking into account the particular, sometimes complex ways in which programs work under a variety of circumstances.

5.2 Beyond the RCT—what counts as evidence?

Alternatively, we might recognise different kinds of evidence beyond that offered by the RCT, giving equal weight to other rigorous research methods, arguably better aligned with real-world educational contexts (Parra & Edwards, 2024). Thomas (2016, p1) noted that, ‘by elevating the status of a particular form of inquiry, we frame questions and answers about education in such a way that research is constrained and stunted’ and argues for ‘a restoration of respect for the heterogeneity of education inquiry – a matrix involving different methods or combinations of methods assessed for their suitability for particular projects in addressing the multifactorial issues that abound in the messy worlds they study’.

In the following section, we will provide examples of studies from genre theory and systemic functional linguistics as evidence of what might be overlooked, when rich and potentially valuable aspects of research work are at risk of being traded for an emphasis on a specific approach such as experimentation.

6 A sampling of GT/SFL contributions

Returning to the article by Green and others, the issue of what counts as evidence raises the question of the 7846 peer-reviewed studies that were excluded from consideration. To this end, we present an overview of a small selection of studies that have served to advance language and literacy educational practice in a variety of ways in Australia and internationally. These apparently disparate studies have in common their use of genre theory and systemic functional linguistics as important informing theories. In the first set of studies we present, SFL was used as a fine-grained analytic tool to track language development from early childhood, across the years of schooling and in different disciplines. In the other examples, SFL was used as a tool to reflect on classroom practice and as a pedagogic resource enabling teachers to be explicit about how language works in the curriculum, contributing to levels of student achievement in literacy and learning.

6.1 Understanding language and literacy development

A key contribution to literacy education by GT/SFL research has been to provide evidence of how language develops from babyhood to adulthood through the analysis of oral and written texts from a meaning-making perspective. Building on Michael Halliday's ground-breaking study of his own child's language (Halliday, 1975), Painter (1998), Torr (2004) and others studied how language develops in toddlerhood and preschool contexts through interaction with caregivers. Moving into the written mode in primary schooling, Martin and Rothery (1980) drew on Halliday's model of language to analyse hundreds of student texts to identify the kinds of social purposes for which the children were writing. They found that predominantly they were making observations about and creating recounts of personal experience. This set in motion a movement investigating the genres that students need to master in response to different pedagogical tasks across the school years.

A major research project, entitled Write it Right (Disadvantaged Schools Program, Metropolitan East Region, 1994), continued this work into the secondary years, with a large team analysing the genres characteristic of the different subject areas, identifying the different stages and typical language features of each text type in achieving their purpose and relating all of these to workplace literacy. Christie and Derewianka undertook more finely grained analyses of some 2000 secondary students' texts in the subjects of English, History and Science funded by the Australian Research Council, resulting in the seminal publication *School Discourse* (2008).

A number of other researchers have collaborated with teachers in investigating the explicit teaching of relevant genres, their structure and language features in classroom contexts across the years of schooling. Jones et al. (2022), for example, examined students' literacy as they transitioned from primary to secondary school. Myhill and Chen (2020) presented an overview of writing development from the early years through to late secondary. Coffin (2006) and Hao and Martin (2024) provided substantial insights into the discourses of History while the language demands of Science have also been particularly well documented (e.g. Halliday & Martin, 1993; Korner et al., 2007; Maton et al., 2021; Martin & Veel, 1998).

Research into reading has also drawn on GT/SFL analysis of text features. Several studies have reported on the role of textual cohesion (Halliday & Hasan, 1976) in reading comprehension (e.g. Carlisle & Rice, 2002; Chapman, 1983; Oakhill & Cain, 2007). Other researchers have investigated the impact of readers' expectations of genre structure on their comprehension of texts (e.g. Schmitz et al., 2017). Martin and Unsworth (2023), Lim (2018), Hsu and Yang (2007) and Buch (2020), among others, have examined pedagogical support for students' interpretation of visual texts and the relationship between visual and verbal elements. And scholars such as Macken-Horarik et al. (2015) and O'Hallaron et al. (2015) have used SFL in their analysis of the development of critical reading skills.

This multi-decade GT/SFL research program has had a significant impact on our understanding of students' literacy development as testified in the host of studies alluded to in the article in question and elsewhere (Jones, [in press](#)). Importantly, such a theoretical framework gives teachers the conceptual tools to transcend the simplistic and linguistically narrow notions of 'sentence structure' and the 'parts of speech', allowing them to focus on meaning-making at different levels (from the whole text through to the paragraph, stage, phase, sentence, group level and below)—and providing them with clear direction on how students can improve their writing, reading and oral language across the years of schooling and across disciplines.

6.2 Longitudinal school studies

The following brief accounts of two longitudinal, single-school studies are included to demonstrate how SFL theory has impacted across the globe: one project is based in the USA and another in Australia. Both studies demonstrate the complex nature of schooling, acknowledging the many factors that have contributed over time to school improvement, with SFL theory playing a part.

The first study took place in the USA, where university researchers collaborated with teachers in a mixed-methods research project to develop a successful school-wide writing program at an urban elementary school with a multilingual school population. The data included professional development records, classroom observations with audio or video recordings, teacher interviews, and students' writing produced over the course of specific genre units, including plans, drafts, and final pieces. Students' performance demonstrably improved over time and within each genre, as evidenced by the state English Language Arts assessment. By the fourth year of intervention, the school had moved from the lowest rating (Level 5) to the top district rating (Level 1), for closing the achievement gap for all students, as well as all subgroups (including high-needs students, low-income students, and English Language Learners) (Brisk & Kaveh, 2020).

While summative test results provided evidence of student progress, they alone did not reveal the many factors that made the progress possible. These factors are identified in the many qualitative research studies published from the project:

- The changes in teachers' knowledge and agency (Brisk, 2021)
- The classroom ecology and practices that led to the success of beginner English language learners (Brisk & Ossa Parra, 2018)
- The struggles and success of students' writing specific genres (Brisk & Alvarado, 2020; Brisk, Tian, & Ballard, 2021)
- The impact on disciplinary teachers embracing language and writing instruction within their content instruction (Hodgson-Drysdale & Rosa, 2015)

These and other detailed findings were garnered from the multiple qualitative studies that emerged from this 10-year SFL genre pedagogy intervention.

The second longitudinal study took place in Australia, a mixed-methods study aimed at evaluating the whole-school implementation of an explicit language-based pedagogy at Pakenham Secondary College in Victoria, Australia (Pomagalska, 2018). Grounded in the functional model of language (Martin, 1999), the comprehensive training was developed and delivered by Lexis Education (Dare & Polias, 2010) to build the capacity of the teachers to implement the pedagogical approach within the whole school and across all curriculum areas.

Analysis of several years of the school's results in national and state assessment regimes, NAPLAN and the Victorian Certificate of Education showed a positive, consistent and significant improvement in literacy and numeracy results. Before the implementation began, the school's NAPLAN results were significantly below the state results. Within two years, many of the measures of literacy and numeracy growth exceeded the state average. Three years into the intervention, all of the measures significantly exceeded the state average.

To complement this analysis of quantitative data, qualitative data was collected using interviews, focus groups and observational data and analysed for content. This

qualitative data was essential to build a comprehensive understanding of how and why this new pedagogy impacted teaching and learning practices.

The analysis of the interviews further supported the above results. There was a consensus among teachers that students produced better texts across all subjects. The improvement was, according to the interviewees, a result of the change in pedagogical practices. This new approach gave teachers a way to understand explicitly how the texts in their subject worked so they felt better prepared and more confident to teach the linguistic features of those texts to their students. Teachers further reported a notable impact on student learning practices and behaviour.

The data also revealed that the implementation was carefully managed over a number of years (Pomagalska, 2018). As with the previously described international study, this complex project would not have been possible without an unswerving leadership that is able to foresee and tackle potential challenges. These included the challenge of changing teachers' practices, ensuring whole school engagement, and the challenges involving maintaining the fidelity of the pedagogical approach advocated by the initial training.

Many factors converged to bring such positive change in this school, including a well-crafted professional development program and the meticulous charting of the implementation guided by a committed and enthusiastic leadership team. And while formal improvement in literacy can be summed up by encouraging numbers, it is often in teachers' voices that we find the core of what such change is about:

I never thought we would get there. I had this little dream that...it was the national average that was hoping we would one day hit, because it's lower than the state average, so I wasn't even thinking of the state, I was trying to get them to the national average...Back in 2013 our kids weren't even in the game, they weren't competing. Now they are matching the state. (Quote from an interview with Literacy Leader in Pomagalska, 2018)

6.3 A study of cross-country implementation of an SFL-underpinned literacy program

The previous two single-school studies used a longitudinal lens on a language-focused pedagogy implementation. In contrast the 'Reading to Learn' project referred directly to a cluster of SFL-based pedagogy initiatives taking place simultaneously across several educational sites. Each site implemented Reading to Learn pedagogy (Rose & Martin, 2012), a program that has been taken up beyond Australia in more than 20 countries around the world and in over 25 languages for more than a decade (e.g. Acevedo et al., 2023).

In 2009, an action research project conducted by the Multilingual Research Institute in Stockholm, Sweden, reported the results of classroom implementation of Reading to Learn pedagogy over a school year (Acevedo, 2010, 2014). This involved a group of 22 teachers from seven primary and secondary schools. Analysis of the pre- and post-reading and writing assessments (Rose & Martin, 2012) showed a distinct reduction in the achievement gap between high and low achieving students. This mirrors the trend in the data collected from a similarly designed R2L project in Australia (Culican, 2005).

The Stockholm project became the catalyst for a 2-year EU-funded international R2L project, Teacher Learning for European Literacy Education (TeL4ELE), conducted in five European countries: Sweden, Denmark, Spain, Portugal and Scotland 2011–2013 (Acevedo, 2023; Acevedo et al., 2016; Coffin et al., 2013). Ninety-eight classroom teachers were trained in the pedagogy in their national languages. The trained teachers implemented

the pedagogy with 2400 students across Europe in seven different languages. The project evaluation of student learning, using pre and post national reading comprehension tests and the R2L writing assessment criteria, found that all groups of students showed improved learning outcomes. The greatest gains were made by the most disadvantaged learners (Coffin et al., 2013, p.12), a finding that also emerged from the data in the earlier Australian and Swedish projects.

Reading to Learn PL has continued in many parts of Sweden. Significantly, educators in Gothenburg have carried out a long-term action research project for over 10 years (2013–2024) (Andersson Varga et al., 2023). The project design was based on the 2010 Stockholm model and used the same data collecting instruments: the DLS reading comprehension test (Järpsten & Taube, 1997) and the R2L writing assessment tool (Rose & Martin, 2012). By 2020, an examination of 8 years of data confirmed that the implementation of the R2L pedagogy had improved learning outcomes for the majority of learners with the lowest achieving students in pre-program assessments making the most significant gains during the implementation year (Andersson Varga et al., 2023; Lövestedt & Acevedo, 2020).

The use of clear, fine-grained criteria based on the functional model of language (Rose & Martin, 2012) provided, and continues to provide, common tools for comparative analyses both within and between the different genre-based classroom initiatives in different contexts and languages (Coffin et al., 2013). The scoring criteria from these analyses were used to generate quantitative data that, when combined with scores from reading assessments, provided credible evidence of literacy achievement at a local level. While claims made by generalising from cases cannot be considered as ‘proof’ in a statistical sense, the derivation and consistency of units of text analysis were clear across the many sites (Freebody, 2003). SFL-guided discourse analysis emerges as a robust framework for ‘analytic generalisability’ (Lincoln & Guba, 2000, p. 112; Yin, 2012, p. 18).

6.4 Classroom discourse studies

The case studies described above provide an overview of the efforts and long-term commitment required for whole-school change. The final example of SFL-based research took place within a much smaller time scale and with a very small number of students.

Scaffolding academic language with educationally marginalised students was a mixed-methods research project supported by the Primary English Teaching Association of Australia (PETAA) in 2017 (Harper & Parkin, 2017). It consisted of case studies in two sites: one in Maningrida, in remote Arnhem Land in Australia, with students who speak up to five languages, including English, and one in suburban Adelaide, in a metropolitan primary school with a high proportion of students from migrant families with English as an Additional Language or Dialect (EALD). One site chose to investigate effective pedagogy for the teaching of Mathematics, the other Science.

The project was an investigation of the moment-by-moment classroom dialogue over 2 weeks in two different school contexts. Using SFL to inform teaching, and as an analytic tool, the teachers and researchers examined the impact of teachers’ language choices on students’ access to powerful language in curriculum tasks. Through pre- and post-tests, student language use was analysed and quantified, using SFL tools for in-depth analysis in both sites of linguistic feature characteristic of academic discourse. The identification of lexical and grammatical language goals for students allowed for the tracking of their

learning through their language use. While of modest scope, the study demonstrated growth in all students' uptake of target language resources, allowing for the development and application of a set of preliminary pedagogic principles which proved robust in the two different sites. These principles then provide guidelines for teachers on how to make systematic and contingently appropriate pedagogic choices at each teaching moment and in highly variable contexts (Harper et al., 2018; Harper & Parkin, 2023; Parkin & Harper, 2020).

6.5 In summary

Each pedagogic intervention reported here puts to the test a highly complex instructional approach incorporating the following:

- Careful lesson planning with a literacy focus
- Respectful acknowledgement of teacher insights
- Intricate relationships between teacher talk, reading and writing activities
- Skilled contingent scaffolding
- Choices in text selection to suit student cohorts across the years of schooling
- Recognition of the literacy demands of different disciplines
- Dynamic theory-building

When a pedagogic approach appears to be effective, but has many ingredients, one of the challenges in figuring out why it 'works' involves careful examination of each component. Even if we know that a particular pedagogic practice 'works', it is difficult to disentangle the elements of the complex teaching and learning endeavour in order to pinpoint which of those elements are the 'effective' one(s). Such a richness of information and insight offered by these instructional studies cannot be easily reduced to quantitative outcomes under highly controlled circumstances. Shanahan (2020) and others (Thomas, 2022; Wyse & Bradbury, 2022) argue that such applied research is essential in translating the findings of basic research for classroom use.

Evaluation at the school and classroom level of approaches emerging from basic or experimental research, and pedagogies for implementing those approaches, needs to be both principled and ethical. To determine whether an approach and its implementation is both appropriate and potentially effective for students in a specific teaching situation, school leaders and classroom teachers will ideally interrogate the approach drawing on well-established research design and evaluation principles (Schwandt et al., 2007; Yin, 2016), using questions such as those proposed by Feez and Cox (2017):

- Can the approach be implemented in a way that ensures students are clear about what is expected of them and why? (transparency)
- Does the benefit of implementing this practice outweigh any burden it might place on teachers and students in terms of time, effort and stress? (practicality)
- Is the evidence on which the claimed benefit of this practice is based meaningful in this specific teaching context, or are the claims unrelated or overstated in relation to the needs of this particular group of students? (validity)
- Will the evidence on which the claimed benefit of this practice is based hold true for this group of students and in this teaching situation at this time? (reliability)
- Does the practice address the real-world demands of the curriculum and the learning needs of this group of students? Does it enable the students to engage meaningfully

with curriculum content and will it help them meet their long-term educational needs and goals? (authenticity)

Such questions, we argue, might be usefully applied to all research findings destined for curriculum and classroom, whether the findings arise from RCTs, meta-analyses or more qualitatively oriented research such as the GT/SFL studies described above.

7 Conclusion

These examples of GT/SFL studies drawn from the over-7000 research projects that are identified and dismissed by Green and colleagues are representative of the myriad of investigations overlooked in favour of a narrow base of evidence for ‘what works’ in literacy education. The insistence on experimental and quasi-experimental research, with RCTs as the sole-gold standard, demonstrates a lack of understanding of the nature of research (McKnight & Morgan, 2022). Policy decisions must be informed by wider criteria than ‘evidence-based’ approaches which use only RCTs. While there is a place for well-designed experimental studies, reliance on oversimplified causal explanations discounts the rich contribution of studies from other research approaches that provide evidence of achievement from different perspectives, recognising the diversity and subtleties of literacy education in classroom settings.

It is understandable but unfortunate that educationists and policymakers reach for the seemingly unequivocal evidence that seems to be provided by experimental research along with the associated meta-analyses and systematic literature reviews that eschew research from other sources, distilling multiple studies of varying quality into a single ‘effect size’. Literacy education is impoverished without teachers’ and decision-makers’ access to more comprehensive and inclusive bodies of theoretically robust and rigorous research evidence.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Not applicable. The paper does not report directly on studies for which funding was received.

Data availability This paper does not draw on primary data but rather on published sources.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval All ethics approvals were received for original studies which are reported on here in their published form.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acevedo, C. (2010). *Will the implementation of reading to learn in Stockholm schools accelerate literacy learning for disadvantaged students and close the achievement gap? A report on school-based action research*. Multilingual Research Institute. <https://www.researchgate.net/publication/355160739>
- Acevedo, C. (2014). Reading to learn: Scaffolding democracy in literacy classrooms in Stockholm. In H. Emery & N. Moore (Eds.), *Teaching, learning and researching reading in EFL* (pp. 298–320). TESOL Arabia Publication.
- Acevedo, C. (2023). A story of international cooperation: teacher learning for European literacy education (TeL4ELE). In C. Acevedo, D. Rose, & R. Whittaker (Eds.), *Reading to learn, reading the world: How genre-based literacy pedagogy is democratizing education* (pp. 144–160). Equinox.
- Acevedo, C., Coffin, C., Gouveia, C., Löfstedt, A. C., & Whittaker, R. (2016). Teacher learning for European literacy education (TeL4ELE) 2011–2013 project. In J. de Silva & S. Feez (Eds.), *Exploring literacies: Theory, research and practice* (pp. 327–333). Palgrave Macmillan.
- Acevedo, C., Rose, D., & Whittaker, R. (Eds.). (2023). *Reading to Learn, reading the world: How genre-based literacy pedagogy is democratizing education*. Equinox.
- Anderson Varga, P., Mitiche, A., Sandberg, J., & Staf, S. (2023). Scaffolding a long-term professional development project in a disadvantaged Swedish context. In C. Acevedo, D. Rose, & R. Whittaker (Eds.), *Reading to Learn, reading the world: How genre-based literacy pedagogy is democratizing education* (pp. 126–143). Equinox.
- Australian Educational Research Organisation (AERO). (2023). *AERO's submission to the Review to inform a better and fairer education system*. <https://www.edresearch.edu.au/other/submissions/review-inform-better-fairer-system>
- Barrenechea, I., Beech, J., & Rivas, A. (2023). How can education systems improve? A systematic literature review. *Journal of Educational Change*, 24(3), 479–499. <https://doi.org/10.1007/s10833-022-09453-7>
- Biesta, G. J. J. (2010). Why ‘What Works’ still won’t work: From evidence-based education to value-based education. *Studies in Philosophy and Education*, 29(5), 491–503. <https://doi.org/10.1007/s11217-010-9191-x>
- Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine* (1982), 75(12), 2299–2306. <https://doi.org/10.1016/j.socscimed.2012.08.032>
- Brisk, M. E. (2021). Theory inspired best practices: Elementary teachers appropriate SFL theory to inform their practice. In M. E. Brisk & M. J. Schleppegrell (Eds.), *Language in action: SFL theory across contexts* (pp. 13–32). Equinox.
- Brisk, M. E., & Alvarado, J. (2020). Uncovering ‘The Story’ behind meaningful texts: Bilingual students’ intentions and linguistic choices. In M. Zappavigna & S. Dreyfus (Eds.), *Discourses of hope and reconciliation: On J. R. Martin's contribution to systemic functional linguistics* (pp. 167–184). Bloomsbury Publishing.
- Brisk, M. E., & Kaveh, Y. M. (2020). Mainstream teachers for successful multilingual classrooms: The case of a school that embraced a genre-based pedagogy to teach writing. In S. Hammer, K. M. Viesca, & N. L. Commins (Eds.), *Teaching content and language in the multilingual classroom: International research on policy, perspectives, preparation and practice* (pp. 145–167). Routledge. <https://doi.org/10.4324/9780429459443-12>
- Brisk, M. E., & Ossa Parra, M. (2018). Mainstream classrooms as engaging spaces for emergent bilinguals: SFL theory, catalyst for change. In R. Harman (Ed.), *Bilingual learners and social equity: Critical approaches to systemic functional linguistics* (pp. 127–151). Springer International Publishing. https://doi.org/10.1007/978-3-319-60953-9_7
- Brisk, M. E., Tian, Z., & Ballard, E. (2021). Autobiography writing instruction: The journey of a teacher participating in a systemic functional linguistics genre pedagogy professional development. *System (Linköping)*, 97, 102429. <https://doi.org/10.1016/j.system.2020.102429>
- Buch, B. (2020). Using SFL as a powerful tool for analyzing multimodal texts from vocational education. *Globe: A Journal of Language, Culture and Communication*, 10, 67–81. <https://journals.aau.dk/index.php/globe/article/view/5879/5178>
- Carlisle, J., & Rice, M. (2002). *Improving reading comprehension research-based principles and practices*. York Press.
- Carolan, J., & Zielezinski, M. (2019). Debunking the ‘Gold Standard’ myths in edtech efficacy. *Edsurge*. <https://www.edsurge.com/news/2019-05-21-debunking-the-gold-standard-myths-in-edtech-efficacy>
- Cartwright, N. (2009). What are randomised controlled trials good for? *Philosophical Studies*, 147(10), 59–70.
- Chapman, J. (1983). *Reading development and cohesion*. Pearson Education.

- Christie, F., & Derewianka, B. (2008). *School discourse : Learning to write across the years of schooling*. Continuum.
- Coffin, C. (2006). *Historical discourse : The language of time, cause, and evaluation*. Continuum.
- Coffin, C., Acevedo, C., & Löwstedt, S. (2013). *Teacher learning for European literacy education (TeL-4ELE) final report public part*. <https://doi.org/10.13140/RG.2.2.20401.30565> <https://docta.ucm.es/entities/publication/b1eed2ae-3293-4c2d-9748-1a10b5694056>
- Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980-2016. *Educational Research (Windsor)*, 60(3), 276–291. <https://doi.org/10.1080/00131881.2018.1493353>
- Culican, S. J. (2005). *Learning to read, reading to learn: A middle years literacy intervention research project, Final report 2003-4*. <https://doi.org/10.13140/RG.2.2.13690.41921>
- Dare, B., & Polias, J. (2010). *Literacy for learning*. South Australian Dept for Education and Child Development.
- Disadvantaged Schools Program, Metropolitan East Region. (1994). *Write it right: Resources for literacy and learning*. Disadvantaged Schools Program, Metropolitan East Region, NSW Dept of School Education.
- Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries?: An analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, 31(2), 242–270. <https://doi.org/10.1093/wbro/kw004>
- Feez, S., & Cox, R. (2017). *Understanding research and evidence: A guide for teachers*. Primary English Teaching Association Australia (PETAA) Paper No. 209. <https://petaa.edu.au/Mp/Mp/Curric-Res/PPs/PETAA-PAPER-209.aspx>
- Freebody, P. (2003). *Qualitative research in education: Interaction and practice*. SAGE.
- Gale, T. (2017). What's not to like about RCTs in education? In A. Childs & I. Menter (Eds.), *Mobilising teacher researchers: Challenging educational inequality* (pp. 207–223). Routledge. <https://doi.org/10.4324/9781315160320>
- Ginsburg, A., & Smith, M. S. (2016). *Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the what works clearinghouse*. <https://www.carnegiefoundation.org/resources/publications/do-randomized-controlled-trials-meet-the-gold-standard>
- Green, C., Giblin, I., & Mulder, J. (2024). A systematic narrative synthesis review of the effectiveness of genre theory and systemic functional linguistics for improving reading and writing outcomes within K-10 education. *The Australian Journal of Language and Literacy*, 47(2), 203–223. <https://doi.org/10.1007/s44020-024-00060-y>
- Halliday, M. A. K. (1975). *Learning how to mean: Explorations in the functions of language*. Edward Arnold.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
- Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Falmer Press.
- Hao, J., & Martin, J. R. (2024). *The discourse of history: A systemic functional linguistics perspective*. Cambridge University Press.
- Harper, H., & Parkin, B. (2017). *Scaffolding academic language with educationally marginalised students*. Primary English Teaching Association Australia (PETAA). <https://www.researchgate.net>
- Harper, H., & Parkin, B. (2023). Scaffolding dialogue with marginalised students in the middle years. In A. Thwaite, A. Simpson, & P. T. Jones (Eds.), *Dialogic pedagogy: Discourse in contexts from pre-school to university* (pp. 109–123). Routledge. <https://doi.org/10.4324/9781003296744>
- Harper, H., Lotherington, M., & Parkin, B. (2018). Carrying the conversation in my head: Classroom dialogue in a remote Aboriginal setting. In P. T. Jones, A. Simpson, & A. Thwaite (Eds.), *Talking the talk: Snapshots from Australian classrooms* (pp. 75–87). Primary English Teaching Association Australia (PETAA).
- Helmer, J., Harper, H., Lea, T., Wolgemuth, J. R., & Chalkiti, K. (2013). Challenges of conducting systematic research in Australia's Northern Territory. *Asia Pacific Journal of Education*, 34(1), 36–48. <https://doi.org/10.1080/02188791.2013.809692>
- Henneberger, A. K., Rose, B. A., Feng, Y., Johnson, T., Register, B., Stapleton, L. M., Sweet, T., & Woolley, M. E. (2023). Estimating student attrition in school-based prevention studies: Guidance from state longitudinal data in Maryland. *Prevention Science*, 24(5), 1035–1045. <https://doi.org/10.1007/s11121-023-01533-1>
- Hodgson-Drysdale, T., & Rosa, H. (2015). Go with the flow. *Science and Children*, 52(6), 32–38. <https://go.openathens.net/redirector/unisa.edu.au>
- Hsu, P., & Yang, W. (2007). Print and image integration of science texts and reading comprehension: A systemic functional linguistics perspective. *International Journal of Science and Mathematics Education*, 5, 639Y659.

- Hubbard, R., Haig, B. D., & Parsa, R. A. (2019). The limited role of formal statistical inference in scientific inference. *The American Statistician*, 73(1), 91–98. <https://doi.org/10.1080/00031305.2018.1464947>
- Hunter, J. (2022). *Improving student learning in Australia: Submission to the Review of the National School Reform Agreement*. Grattan Institute. https://grattan.edu.au/wp-content/uploads/2022/06/Grattan-Institute_Productivity-Commission_NSRA_Submission_20220617_Final.pdf
- Järpsten, B., & Taube, K. (1997). *Word knowledge test for grades 4-6*. Psychology Publishing House.
- Jones, P. T. (In press). Language in education. In D. Butt, F. Navarro, C. Webster, & J. Webster (Eds.), *The Bloomsbury companion to systemic functional linguistics* (2nd ed.). Bloomsbury.
- Jones, P. T., Matrugiolo, E., & Edwards-Groves, C. (Eds.). (2022). *Transition and continuity in school literacy development*. Bloomsbury.
- Joyce, K. E., & Cartwright, N. (2018). Meeting our standards for educational justice: Doing our best with the evidence. *Theory and Research in Education*, 16(1), 3–22. <https://doi.org/10.1177/1477878518756565>
- Kaplan, A., Cromley, J., Perez, T., Dai, T., Mara, K., & Balsai, M. (2020). The role of context in educational RCT findings: A call to redefine “Evidence-Based Practice”. *Educational Researcher*, 49(4), 285–288. <https://doi.org/10.3102/0013189X20921862>
- Korner, H., McInnes, D., & Rose, D. (2007). *Science literacy*. NSW Adult Migrant Education Service (AMES).
- Koutsouris, G., & Norwich, B. (2018). What exactly do RCT findings tell us in education research? *British Educational Research Journal*, 44(6), 939–959. <https://doi.org/10.1002/berj.3464>
- Lim, F. V. (2018). Developing a systemic functional approach to teach multimodal literacy. *Functional Linguistics*, 5(13). <https://functionallinguistics.springeropen.com/articles/10.1186/s40554-018-0066-8#citeas>
- Lincoln, Y., & Guba, E. (2000). The only generalization is there is no generalization. In R. Gomm, M. Hammersley, & P. Foster (Eds.), *Case study method: Key issues, key texts*. Sage.
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>
- Lövstedt, A.-C., & Acevedo, C. (2020). *Rapport reading to learn 2013-2020*. Göteborgs Stad. https://goteborg.se/wps/wcm/connect/d1a91de6-1862-4242-853f-d7f5772f21e6/R2L-rapport_2013-2020.pdf?MOD=AJPERES
- Macken-Horarik, M., Sandiford, C., Love, K., & Unsworth, L. (2015). New ways of working ‘with grammar in mind’ in school English: Insights from systemic functional grammatics. *Linguistics and Education*, 31, 145–158. <https://doi.org/10.1016/j.linged.2015.07.004>
- Malina, M., Hanne, S., Nørreklit, F., & Selto, H. (2010). *Lessons learned: Advantages and disadvantages of mixed method research*. University of Colorado.
- Martin, J. R. (1999). Mentoring semogenesis: ‘Genre-based’ literacy pedagogy. In F. Christie (Ed.), *Pedagogy and the shaping of consciousness linguistic and social processes* (pp. 123–155). Continuum.
- Martin, J. R., & Rothery, J. (1980). *Working papers in linguistics No. 1*. University of Sydney <https://search.worldcat.org/title/Working-papers-in-linguistics-No.-1-Writing-project-report-1980-J.R.-Martin-Joan-Rothery/oclc/222951731>
- Martin, J. R., & Unsworth, L. (2023). Reading images for knowledge building: Analysing infographics in school science. *Routledge*. <https://doi.org/10.4324/9781003164586>
- Martin, J. R., & Veel, R. (1998). *Reading science: Critical and functional perspectives on discourses of science*. Routledge.
- Maton, K., Martin, J. R., & Doran, Y. J. (Eds.). (2021). *Teaching science: Language, knowledge and pedagogy*. Routledge.
- McKnight, L., & Morgan, A. (2022). Scientific measurement won’t answer all questions in education. We need teacher and student voices, too. *The Conversation* (March 9, 2022). <https://theconversation.com/scientific-measurement-wont-answer-all-questions-in-education-we-need-teacher-and-student-voices-too-178167>
- Morrison, K., & van der Werf, G. (2012). Editorial. *Educational Research and Evaluation*, 18(4), 303–306. <https://doi.org/10.1080/13803611.2012.689168>
- Mosteller, F., & Boruch, R. F. (2002). *Evidence matters: Randomized trials in education research*. Brookings Institution Press.
- Muller, S. M. (2020). The implications of a fundamental contradiction in advocating randomized trials for policy. *World Development*, 127, 104831. <https://doi.org/10.1016/j.worlddev.2019>
- Myhill, D., & Chen, H. (2020). Developing writers in primary and secondary school years. In H. Chen, D. Myhill, & H. Lewis (Eds.), *Developing writers across the primary and secondary years* (pp. 1–18). Routledge. <https://doi.org/10.4324/9781003018858-1>

- O'Hallaron, C. L., Palincsar, A. S., & Schleppegrell, M. J. (2015). Reading science: Using systemic functional linguistics to support critical language awareness. *Linguistics and Education*, 32(Part A), 55–67. <https://doi.org/10.1016/j.linged.2015.02.002>
- Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies* (pp. 47–71). Psychology Press. <https://doi.org/10.4324/9780203810033>
- Painter, C. (1998). Preparing for school: Developing a semantic style for educational knowledge. In F. Christie (Ed.), *Pedagogy and the shaping of consciousness: Linguistic and social processes* (pp. 66–87). Cassell.
- Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40(2), 184–202. <https://doi.org/10.1598/RRQ.40.2.3>
- Parkin, B., & Harper, H. (2020). The application of three exotropic theories in education. *Mind, Culture, and Activity*, 27(3), 233–248. <https://doi.org/10.1080/10749039.2020.1794008>
- Parra, J. D., & Edwards, D. B. (2024). Challenging the gold standard consensus: Randomised controlled trials (RCTs) and their pitfalls in evidence-based education. *Critical Studies in Education*, 1–18. <https://doi.org/10.1080/17508487.2024.2314118>
- Pawson, R., & Tilley, N. (2004). *Realist evaluation*. Sage http://www.communitymatters.com.au/RE_chapter.pdf
- Phillips, D. C. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. *Teachers College Record* (1970), 109(13), 376–402. <https://doi.org/10.1177/016146810710901315>
- Pomagalska, D. (2018). *Evaluation of the whole-school implementation of a language-based pedagogy at Pakenham Secondary College*. <https://lexised.com/report-ssc/>
- Rose, D., & Martin, J. R. (2012). *Learning to write, reading to learn: Genre, knowledge and pedagogy in the Sydney school*. Equinox.
- Schmitz, A., Gräsel, C., & Rothstein, B. (2017). Students' genre expectations and the effects of text cohesion on reading comprehension. *Reading & Writing*, 30(5), 1115–1135. <https://doi.org/10.1007/s11145-016-9714-0>
- Schwandt, T. A., Lincoln, Y. S., & Guba, E. G. (2007). Judging interpretations: But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Evaluation*, 11–25.
- Scriven, M. (2008). A summative evaluation of RCT Methodology: An alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, 5(9), 11–24. <https://doi.org/10.56645/jmde.v5i9.160>
- Sedgwick, P. (2015). Randomised controlled trials: Understanding confounding. *BMJ (Online)*, 351, h5119–h5119. <https://doi.org/10.1136/bmj.h5119>
- Shanahan, T. (2020). What constitutes a science of reading instruction? *Reading Research Quarterly*, 55(51), S235–S247.
- Shelley, M., Yore, L. & Hand, B. (2009). *Quality standards in literacy and science education: International perspectives and gold standards*. https://www.researchgate.net/publication/297226191_Quality_Research_in_Literacy_and_Science_Education_International_Perspectives_and_Gold_Standards
- Simpson, A. (2019). The evidential basis of "evidence-based education": An introduction to the special issue. *Educational Research and Evaluation*, 25(1-2), 1–6. <https://doi.org/10.1080/13803611.2019.1617979>
- Singh, P. (2002). Pedagogising knowledge: Bernstein's theory of the pedagogic device. *British Journal of Sociology of Education*, 23(4), 571–582. <https://doi.org/10.2307/1393313>
- Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education*, 3(3), 285–289. <https://doi.org/10.4300/JGME-D-11-00147.1>
- Thomas, G. (2016). After the gold rush: Questioning the "Gold Standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review*, 86(3), 390–411. <https://doi.org/10.17763/1943-5045-86.3.390>
- Thomas, P. L. (2022). The Science of Reading movement: The never-ending debate and the need for a different approach to reading instruction. *National Education Policy Center*. <http://nepc.colorado.edu/publication/science-of-reading>
- Tikly, L. (2015). What works, for whom, and in what circumstances? Towards a critical realist understanding of learning in international and comparative education. *International Journal of Educational Development*, 40, 237–249. <https://doi.org/10.1016/j.ijedudev.2014.11.008>
- Torr, J. (2004). Talking about picture books: The influence of maternal education on four-year-old children's talk with mothers and pre-school teachers. *Journal of Early Childhood Literacy*, 4(2), 181–210. <https://doi.org/10.1177/1468798404044515>

- Troyer, M. (2022). The gold standard for whom? Schools' experiences participating in a randomised controlled trial. *Journal of Research in Reading*, 45(3), 406–424. <https://doi.org/10.1111/1467-9817.12395>
- Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6), 3045–3089. <https://doi.org/10.1093/jeea/jvaa019>
- Wrigley, T., & McCusker, S. (2019). Evidence-based teaching: A simple view of "science". *Educational Research and Evaluation*, 25(1-2), 110–126. <https://doi.org/10.1080/13803611.2019.1617992>
- Wyse, D., & Bradbury, A. (2022). Reading wars or reading reconciliation? A critical examination of robust research evidence, curriculum policy and teachers' practices for teaching phonics and reading. *Review of Education*. <https://doi.org/10.1002/rev3.3314>
- Yin, R. K. (2012). *Applications of case study research* (3rd ed.). Sage Publications.
- Yin, R. K. (2016). *Qualitative research from start to finish* (2nd ed.). Guilford Press.