Article

Education of Students with Disabilities, Science, and Randomized Controlled Trials

Research and Practice for Persons with Severe Disabilities
2021, Vol. 46(3) 132–145

DOI: 10.1177/15407969211032341
rpsd.sagepub.com


Samuel L. Odom[1,2]

Abstract

The purpose of this article is to examine the application of randomized controlled trial (RCT) methodology for determining the efficacy of school-based interventions in general and special education. In education science, RCTs are widely acknowledged as the gold standard of efficacy research, with other methodologies relegated to a lower level of credibility. However, scholars from different disciplines have raised a variety of issues with RCT methodology, such as the utility of random assignment, external validity, and the challenges of applying the methodology for assessing complex service interventions, which are necessary for many students with disabilities. Also, scholars have noted that school-based RCT studies have largely generated low effect sizes, which indicate that the outcomes of the interventions do not differ substantially from services as usual. The criticisms of RCT studies as the primary methodology in school-based intervention research for students with disabilities are offered along with recommendations for extending the acceptability of a broader variety of research approaches.

Keywords

science, education, RCT

Most reasonable people place their faith in science. Originating with Francis Bacon in the 16th century as a way of understanding how the world works (Schwarz, 2014), the scientific method has guided the great discoveries of our time (e.g., polio vaccine), and at this writing, there is hope it will lead us out of a pandemic. For much of its history, the fields of education and special education for students with disabilities were "science poor." Currently, there is great interest in discovering interventions or programs, which occasionally are multicomponents that produce specific outcomes for individual students with disabilities when employed in specific contexts (Frey et al., 2005). Striving to produce such evidence, the mainstream of education, psychology, health sciences, and other disciplines has largely adopted the randomized controlled trial (RCT) as the methodological gold standard for providing the evidence for evidence-based practice. Although can provide evidence of the efficacy of instructional and intervention practices, turning to it as the single methodology of choice has relegated other scientific methodologies to a lower level of credibility. In this article, I describe features of RCTs, briefly review their historical roots, and describe the emergence of RCTs as a source for scientific knowledge in the 21st century. The challenges of employing RCTs in authentic school settings will be highlighted, which will draw into question the appropriateness of RCTs as the gold standard for educational research for individuals with disabilities. The article will conclude with potential alternatives.

[1]The University of North Carolina at Chapel Hill, USA
[2]San Diego State University, CA, USA

Corresponding Author:
Samuel L. Odom, 7615 Jennite Drive, San Diego, CA 92119, USA.
Email: slodom@unc.edu

Definition of an RCT

In special education research and other disciplines, the RCT is an experimental method for detecting the effects of an intervention, instructional practice, curriculum, or other types of programs that might benefit children, youth, and adults with disabilities. Also known in education as a two-group, pretest–posttest control group design (Campbell & Stanley, 1963),[1] RCTs require researchers to randomly assign participants to experimental groups. The assumption is that random assignment will result in groups that are equivalent on variables of interest at the beginning of the study. These variables of interest (e.g., IQ, adaptive behavior, communication skills, self-determination) may be the outcome measures (i.e., dependent variables) in the study or characteristics of the participants (e.g., diagnosis or special education classification, biological sex, socioeconomic status). In most current applications of RCTs, experimenters assess group characteristics at the beginning of the study to ensure the groups are equivalent, although there are post-test only designs that assume without measurement that random assignment produces equivalence (Shadish et al., 2002). A basic assumption of random assignment is that it produces equivalence on potentially confounding variables that are not measured or may even be unknown.

After random assignment and pretest, measures have been collected, experimenters usually implement an intervention/program with one group while the members of the other group receive usual practices or engage in a "placebo" or contact-control intervention/program. At the conclusion of the delivery of the program, the experimenters assess the dependent variables again and analyze average differences between the groups that did and did not receive the intervention. Because random assignment occurred, the experimenter infers that the intervention *caused* the differences between the two groups.

RCTs and Field Experimentation

Although features of RCTs (e.g., contrasting group outcomes, blinded conditions) had been used in medical research earlier in the 1900s, numerous methodologists (e.g., Bothwell et al., 2016; Deaton, 2020) attribute the introduction of RCTs into medical research to Austin Bradford Hill in England, whose work began before World War II. Promotion of RCTs as the most valid scientific evidence of effective health practice resulted from Archie Cochrane's work and his influential book, *Effectiveness and Efficiency*, published in 1972. This work led to the development of the evidence-based medicine movement, propelled by Sackett and colleagues (1996). Methodologists acknowledge that the RCT was originally developed for laboratory research (Schwarz, 2014), with the most typical applications being pharmaceutical research (Bédécarrats et al., 2017). In fact, when RCTs were adopted for use outside of the laboratory, for example in health care practices, the methodology became known as *field experimentation* (Schwarz, 2014). Kuklick and Kohler (1996) noted field experiments are "a site of compromised work: field sciences have dealt with problems that resist tidy solutions" (p. 1). This designation informally acknowledges the application of RCT experimentation under conditions that cannot be as tightly controlled as would occur in the laboratory.

Since its inception within the medical discipline, RCT methodology has had uptake, although not uncritically, in such diverse fields as sociology (Wahlberg & McGoey, 2007), clinical psychology (Woolfolk, 2015), developmental economics (Bruhn & McKenzie, 2008), criminology (Weisburd et al., 2014), social work and child welfare (Mezey et al., 2015), and education (Styles & Torgerson, 2018). In all of these fields, RCT methodology has been recognized as the "gold standard" for field-based experimental research. In accordance, other forms of research, such as single-case design (SCD), quasi-experimental designs, survey research, econometric research, and qualitative research have become considered as less credible or trustworthy approaches for establishing the effects of practices and programs (Greenhalgh, 2000). McKnight and Morgan (2020) refer to this as the "bullying" effect of RCTs. If not bullying, there is at least a condescension in the field toward non-RCT research (Methods Group of the Campbell Collaboration, 2017).

Recent History of RCTs in Education

In the United States, the National Academy of Sciences convened the Committee on Scientific Principles for Education Research in 2000, with the charge to "review and synthesize . . . the science and practice of

scientific educational research and consider how to support high quality science. . .” (National Research Council, 2001, p. 1). The committee acknowledged that multiple questions exist about the education discipline and different methodologies are important for addressing those questions. However, they gave primacy to the causal question of efficacy and clearly stipulated that RCTs were the only methodology that could establish that causal relationship. This influential report and the co-occurring Education Science Reform Act (2002) led to the formation of the Institute of Education Sciences (IES), with three centers (National Center for Education Evaluation and Regional Assistance [NCEE], National Center for Education Research, and National Center for Special Education Research). These centers have the goal of addressing efficacy and effectiveness questions. In addition, the IES-funded What Works Clearinghouse had the parallel mission of cataloging interventions that proved efficacious. In the United Kingdom, the Department of Education funded a somewhat similar organization, the Education Endowment Foundation (EEF), with an even more focused goal of funding RCT research to examine educational practices and interventions for children from low-income families. Although motivated by specific educational content interests such as literacy interventions, social skills training, positive behavioral intervention and supports, researchers by necessity follow the funding and, when funding is contingent on the employment of a specific experimental design, then increases are seen in its use.

In the discipline of education, the number of studies employing RCT methodologies and associated meta-analyses that only include RCT work has grown dramatically. In a recent systematic review, Connolly et al. (2018) found 1,017 RCT studies in education had been published between 1980 and 2016, with three-fourths published in the last 10 years of the time period covered. The initial increase in published RCTs began around 2003, just about the time that IES began funding RCT research, with another increase around 2011, when EEF also began funding RCT research. With the review ending in 2016 and an accelerating trend, these data likely underestimate the number of RCTs that will be published by the early 2020s.

Are RCTs Worth the Investment?

Has the investment in RCTs in educational research been effective in identifying interventions that work for children and youth? To address this question, Lortie-Forgues and Inglis (2019) conducted a meta-analysis of RCT studies in education funded by the EEF and the NCEE. Across the two organizations, they found 141 distinct trials, with the median number of participants per trial being 2,386. They computed effect sizes that generally analyzed the differences in the mean scores on outcomes between intervention and contrast groups, divided by the standard deviation, which is similar to a Cohen’s *d*. So, a score of 0 indicates no difference between groups, a positive score indicates an effect favoring the groups that received an intervention, and a negative score indicates that the control group scored higher. In education, a rule of thumb for RCT studies is that they may be expected to generate small to moderate effect size ranges from .25 to .50 (Lipsey et al., 2012). Effect size estimates ranged from −0.16 to 0.74, with a median of 0.03. The unweighted mean of the effect size estimates was 0.06, 95% CI [0.04, 0.08]. The authors also computed a Bayesian analysis, which quantified the relative evidence that the data provide for one hypothesis compared with another (i.e., indicated how likely it was that the study confirmed an alternative [to the null] hypothesis, confirmed the null hypothesis, or was uninformative [i.e., not sufficiently powered to detect a difference]). They noted that confirming a null hypothesis is informative because it indicates what does not work. They found that 23% of the studies confirmed a positive effect associated with the intervention of interest, 38% confirmed an effect in favor of the control/contrast group, and 40% were uninformative (lacked the power to detect and effect). The first finding is consistent with a keynote speech by the Director of IES at the 2019 Project Directors Meeting (Schneider, 2019), indicating that approximately 75% of the studies funded by IES were not finding significantly positive effects for the interventions investigated.

From these data, one could conclude that the educational interventions examined by RCTs are just not effective, although this would not explain the uninformative nature of 40% of the studies examined or the nearly 40% of the studies finding counterfactual effects that Lortie-Forgues and Inglis identified. An alternative conclusion could be the gold standard methodology adopted for educational research is indeed tarnished, which is a conclusion being drawn by investigators from other disciplines (Bédécarrats et al., 2017;

Cartwright, 2007; Deaton, 2020). In the field of special education, as well as the educational discipline in general, a number of the required features of RCTs are challenging to implement. Those challenges remove the luster and reduce the carat quality of many RCT studies in education and special education.

School-Based Interventions and Complexity

In laboratory-based RCTs, researchers attempt to control all variables that could affect the outcome except for the treatment administered (the independent variable). In applying laboratory-based RCT methodology in field settings, there is an assumption of sufficient experimental control. Investigators in the health care field (Bonell et al., 2012), with increasing recognition in other disciplines (Wahlberg & McGoey, 2007), have acknowledged the complexity of service environments and the challenges experimenters encounter when they design intervention programs with proper controls (i.e., ensure that they have isolated the independent variable).

The concept of complex services intervention (CSI) has been prominent in the health care discipline for at least two decades. Fenton (2000) drew a distinction between "treatments where a therapeutic agent . . . can be specified with precision and less readily defined psychosocial and service system interventions" (p. 113). In applying this concept to medical education, McGaghie (2011) provided the example of a program that promoted scientifically based surgical procedures through simulation-based medical education, as a CSI. This approach would qualify in that it is a multicomponent program implemented in a complex social system (e.g., hospital, different health care providers). In health care, an example of an intervention that is less complex would be delivery of flu immunization in a health care clinic.

Other scholars have noted similar complex service models occurring in mental health, public health, social services, and criminal justice fields (Campbell et al., 2000; Proctor et al., 2009; Wolff, 2000). Characteristics of such CSIs are varieties of staffing arrangements in the contexts in which they are implemented, variations in motivation in the recipients of the service interventions and service deliverers, the necessity of delivery of intervention in a sequenced way across a lengthy time period, implementation chains that are occasionally nonlinear, variation in the power of individual components due to the influence of the context, and interventions that feed back on themselves (e.g., they change the conditions in which they were initially implemented; McGaghie, 2011). However, the reason that such interventions are developed (rather than a singular treatment in a clinic) is because the needs of the patient and the contexts in which they receive services are multidimensional.

Schools, for which educational interventions are designed and in which they are implemented, are themselves complex organizations. When implemented in well-defined settings in a school (e.g., a special or general education classroom, a well-specified class period), interventions may be feasibly evaluated by randomly assigning students or classrooms to treatment and control conditions. However, intervention programs designed to meet the needs of students with disabilities in schools need to be comprehensive, are often multi-component, and by law have to be individualized (i.e., which in the medical community would be called personalized). The multicomponent nature of the intervention, sometimes consisting of different component features to match the needs of individual students and implemented by different individuals in different school settings, qualifies these programs as CSIs. As an example, Odom and colleagues (2014) designed a school-based comprehensive treatment program for high school students with autism. The program required the formation of an autism team in the school, assessment of program quality, an action plan to address quality issues, and delivery of nine intervention components with students having different ability levels and learning needs. It was a program with many moving parts and complex in its delivery over a 2-year period. Yet, in order to meet the diverse needs of students with autism in typical high schools, all of the "moving parts" were essential (Steinbrenner et al., 2020).

Questions arise about the appropriateness of RCTs for CSIs. Marchal et al. (2013) have characterized RCTs and complex interventions as oxymorons. They noted that the challenges of employing RCTs with complex intervention are (a) nonlinearity of effects, (b) the necessity of adaptation of an intervention for individual contexts (which may vary considerably), (c) previous history of the organization, and (d) human agency (i.e., acceptance or resistance to the intervention). Added to this list for public schools is the frequent

turnover in staff. In response to criticism of the quality of educational research not adhering to the rigorous quality of the hard sciences, Berliner (2002) noted that education science was the "hardest-to-do" of sciences. He emphasized that education scientists conduct their research in conditions that laboratory scientists would find intolerable. His statements reflect the concerns of many school-based researchers who attempt to employ a methodology originally developed for a laboratory setting and who constantly face the challenges of the complex features of the school context and the intervention employed.

Assumption of Group Equivalence and the Law of Large Numbers

As noted previously, a basic tenet of RCTs is that equivalence between groups will be established best through random assignment (Bruhn & McKenzie, 2008). Such equivalence is necessary to avoid biasing the assessment of an intervention's effect in one direction or another. As noted, the major assumption is that such randomization will control or minimize differences on variables that are not measured or may not be known. For example, in a study of a work-based learning (WBL) program, if more students with intellectual disability who had higher IQ scores were assigned to the intervention group, there is a possibility that they might perform better on the WBL tasks at posttest because of the initial differences in IQ scores between the two groups. The "law of large numbers" comes into play here. The success of random assignment in achieving group equivalence is affected by the number of participants in the study. In their hypothetical examination of random assignment of patients with a disease that occurs in 15% of the population, Kernan et al. (1999) found that nonequivalent groups (i.e., differing 10% or more on disease variable) occurred 33% of the time when the *n* was 30, 24% when the *n* was 50, 10% when the *n* was 100, and 3% when the *n* was 300.

Obeying the law of large numbers is particularly pertinent for educational studies in which participants have "low-prevalence" conditions (e.g., some students with severe disabilities). Often students are enrolled in a single class, randomly assigning students to conditions within a class is problematic because a teacher may have difficulty in using the experimental intervention or approach with some students (in the experimental group) and not others (in the control group). Similarly, when classes within schools are randomly assigned to conditions, teachers in the experimental group may share information about the treatment with teachers in the control group who then may use it with their students. In both conditions, there is a risk of "contamination" of the independent variable. As such, schools must then become the unit for randomization and a multilevel analytic design may be employed to control for "nesting" effects within schools. This may also be a problem when students with certain characteristics, such as severe disabilities, are the participants in the study and there are only a relatively small number in any one school. Although randomization at the school level may control for treatment contamination, it may well increase the complexity of implementation and the cost of the study.

An approach for promoting group equivalence in small samples has been to match or stratify samples before randomly assigning participants to conditions. For example, students might be matched on special education classification (e.g., intellectual disability) or demographic characteristics (e.g., SES), or schools might be matched by school district before they are randomly assigned to groups. Such matching introduces some subjectivity into the randomization process (i.e., the researcher decides on which variable to match), which is counter to the objectivity of pure random assignment (Bédécarrats et al., 2017). There is also the assumption that the matching variable will be associated with the unmeasured variables that could affect outcomes, although this is an untestable and, to some extent, a faith-based assumption. Another strategy is to repeat the random assignment (i.e., re-randomize) when nonequivalent groups occur in the initial random assignment (Bruhn & McKenzie, 2008) until equivalence between groups is achieved. Again, this then introduces the experimenter into the randomization process, which random assignment was designed to avoid.

Although a necessary component of RCTs, random assignment may create logistical issues. In educational research, there are rarely single-blind studies in which the participants do not know which intervention condition they are receiving. In fact, institutional review boards require that participants be knowledgeable about the treatment they may potentially receive (McPherson et al., 2020). Researchers

have discussed the ethics of random assignment, in which some participants receive a benefit from a treatment and others do not, which is sometimes addressed by having waitlist controls. Researchers correctly note that the effects of an intervention are not known until after the study is completed, so the research ethics concern of such assignment may be assuaged. However, researchers do generally begin with the hypothesis that the intervention will have a positive effect (compared with the control) and have to communicate the value of this effect to participants through informed consent. When teachers, students and their parents, are randomly assigned to control groups after they have agreed to participate in an experimental intervention that could have some value to them, reactive effects may occur (e.g., attrition in the control group, greater effort by teachers in the control group). In the RCT studies funded through EEF, Edovald and Nevill (2021) reported sometimes reactive effects of participants assigned to services-as-usual (SAU) groups (e.g., high attrition). Anecdotally, in a recent RCT study that stratified random assignment (i.e., citation withheld by author for confidentiality), two schools within a district were randomly assigned to a school-wide high school intervention and a SAU condition. Although notified ahead of time about the random assignment process, the parents and staff of the school assigned to the SAU condition attributed the assignment to the district administration and implied that they might take legal action related to the assignment.

External Validity

Although RCTs establish a high methodological bar for random assignment, measurement, and analysis, the findings for a single RCT study are context-bound. That is, a researcher may be able to say that an intervention produced certain positive outcomes, but methodologists across fields agree that the findings only apply to the participants, classes, and or schools in the studies (Bothwell et al., 2016; Joyce & Cartwright, 2020). Deaton (2020) has noted for some researchers there is a "contrast between the care that goes into running an RCT and the carelessness that goes into advocating the use of its results" (p. 11). The current practice of including consort charts to display the participant selection process has allowed researchers to describe the sample recruited, the number of participants approached but declined to participate, and the randomization. Essentially, it provides information about the context to which the findings are bound. Although external validity is a major criticism of SCD research by group design advocates, relegating it to a lower level of experimental quality, the context-bound nature of results applies to both SCD and RCT studies.

So, how does one build evidence that an instructional program or intervention has efficacy and is effective? Such a process requires replication, perhaps using the same methodology or potentially different methodologies. Also, independent replication and aggregation of studies by different research groups build confidence that an intervention can actually be implemented in a reliable way and produce similar outcomes (Bédécarrats et al., 2017; Siddiqui et al., 2018). Cartwright (2019) proposes that to have confidence in the effectiveness of an intervention, one has to build "a bird's nest" of inter-related ideas, results, analyses, and reasoning. RCTs can be a part of such a bird's nest, but their findings do not necessarily trump an aggregation of findings from other methodologies (Hitchcock et al., 2018).

Assessment and Measurement of Dependent Variables

Without reliable and valid assessment of the outcomes linked theoretically or conceptually to the independent variable, it is impossible to determine the effects of an intervention in RCTs or in any quantitative study. Most researchers would acknowledge the importance of unbiased assessments, which is optimally accomplished by assessors who are naïve (blinded) to the experimental conditions. Although blinded assessments are possible in schools, the reactivity of having external assessors may influence practices in the classrooms (e.g., control group teachers may provide more intense instruction when they know a certain type of assessment is occurring) and, thereby, outcomes. But, perhaps the most influential feature of assessment on study outcomes is the overlap between the outcome measures and the intervention. Measures may be viewed as proximal (e.g., project-constructed measures designed to directly measure intervention effects) or distal (e.g., standardized, norm-referenced assessment designed to measure general constructs).

A common experience is that school-based RCTs may affect more proximal measures but fail to produce effects on the standardized, norm-referenced assessments. In their reflection on the RCTs funded by the

EEF, Edovald and Nevill (2021) who were staff with the organization noted the absence of significant effects on standardized measures that were distal to the intervention and more frequent significant effects on proximal measures. Recently, Sam et al. (2021) conducted an efficacy study of an intervention program to promote teachers' use of evidence-based practices with children with autism in 60 elementary schools. They found that program quality increased, teachers increased the number of EBPs used and the fidelity of their implementation, and students accomplished more individualized educational program goals, all in comparison with a SAU group. However, on standardized norm referenced measures (e.g., *Vineland Adaptive Behavior Scales*), they found that children in both groups made gains across the year, but there were no differences between groups. The authors proposed that the special education teachers' instruction was designed to promote students' accomplishment of their individual learning goals. It is likely that a student's goal would overlap with only one or two items on a broader standardized assessment, which was unlikely to have a major impact on the standard score on the distal assessment.

The Nature of the Counterfactual

In most RCTs (i.e., treatment comparison studies being the exception), the control group is considered the "counterfactual." The counterfactual measures "what would have happened to the members subject to the intervention had they not been exposed to it" (European Commission, 2016). In most school-based research, SAU is the counterfactual unless experimenters have constructed a contact control (e.g., a condition comparable in terms of student time and teacher attention but addressing a different outcome). The practices occurring in the SAU setting are of great importance in that they can overlap substantially with interventions being examined. For example, in a comparison of two comprehensive programs for preschool children with autism, Boyd et al. (2014) also included a control condition. The funding agency insisted that the special education control classrooms had to be of high quality. Although not technically SAU (i.e., the quality of the special education programs was higher than would be expected in the community), the special education classes were established as a counterfactual for the two programs being compared. The study, involving a total of 78 classrooms, found that groups of children with autism in all conditions made progress across time, with nonsignificant differences among groups on most measures. However, a careful examination of practices (Hume et al., 2011) found substantial overlap across conditions, which the authors attributed to general high-quality program features. As such, the program quality of the counterfactual obscured differences that might have been observed between the two comprehensive program models and an SAU representative of typical program quality in the community.

The counterfactual may also affect the effect sizes generated by programs. In a professional presentation about the effects of early childhood education for children from low-income circumstances, a nationally recognized authority presented a graph of effect sizes of efficacy studies of programs from the early history of the movement (e.g., Abecedarian Project, Perry Preschool Program) and contemporary programs. The effect sizes had not changed markedly across the generation of programs, and the presenter concluded that as a field we had not made expected progress in improving the impact of these programs. However, across generations, the counterfactual had changed. The control group children in earlier studies often did not have access to early childhood education (i.e., the SAU was no program). In the current generation of programs, the children in the counterfactual grouping almost always have access to some program. For example, in a study of the efficacy of a state prekindergarten program, a finding of no difference was interpreted as state-funded kindergarten having no effects. However, a substantial proportion of the families, when their children were assigned to the SAU condition, enrolled their children in other forms of early childhood education (e.g., Head Start, private preschools, even the state prekindergarten program). The point is that what happens to the control group matters.

The challenge for investigators using RCTs is to be able to measure the features of the counterfactual context that are likely affecting the dependent variables. Fidelity of implementation measures that assess features of the intervention program are one source of information when collected in the control classroom settings (Siddiqui et al., 2018). Those features, however, may be so idiosyncratic that reporting few or none is occurring in the counterfactual setting does not account for other potentially impactful instructional activities that are occurring. For example, in a study of a phonics-based reading intervention program,

designed for struggling readers, Norwich and Koutsouris (2020) reported that phonics-based instruction was also occurring in control classrooms. They also noted that in the control group, teachers' participation in the research project may have heightened their awareness of some students' reading difficulties (e.g., through the informed consent process, reading assessments given by the research team). This heightened awareness may have inadvertently and positively affected the reading instruction they were providing to struggling readers in their classes. Without assessing the features of reading instruction in both groups of classes, it would not be possible to discount this as an influence on the null findings.

Implications and Possible Alternatives

The purpose of this article is not to imply that educational researchers should entirely abandon RCTs, but rather to modify the assumption that RCTs are the gold standard in education research and always the preferred methodology to address educational research questions. An old adage is that: "if you give a toddler a hammer, everything looks like a nail." There is a general perception that if a study does not employ random assignment, then it is not experimental and has a lesser degree of scientific integrity, which flies in the face of the history of experimentation from Francis Bacon until the RCT/evidence-based movement took hold in medicine (Schwarz, 2014). However, criticism of an approach as hegemonic as RCTs use in education requires that alternatives be suggested.

Relevance and Flexibility of RCT Standards

Anyone who has served on a research review committee for agencies that fund educational research would be hard pressed not to see what Cartwright (2007) has called "the vanity of rigor in RCTs" (p. 18). Methodology and methodologists' opinions are often privileged over relevance and educational or social significance. For example, in the IES scoring criteria, it is possible to have a perfect score on the significance and still not be funded if a miscalculated power analysis dropped the overall score on the RCT to slightly below an arbitrary benchmark established by the funding agency. By ratcheting up the rigor of the methodological standards, organizations may have created a situation where 75% of the studies find no effects and average effect sizes are very small. Similarly, Edovald and Nevill (2021) have questioned the sanctity of the arbitrary but inflexible .05 significance level in their discussion of the outcomes of EEF studies. As Cartwright (2019) notes, one could be relatively sure that an effect will occur 9 out of 10 times if a study were done again and still have nonsignificant findings. One recommendation would be to allow some latitude in applying the most rigorous standards of the RCT methodology when other supporting evidence is provided. Such supporting evidence could come from process evaluations.

Process Evaluations

A primary point made about RCT methodology is that it can provide confidence in the effects of an intervention, but such studies may provide very little information about the reason such effects occurred (Bédécarrats et al., 2017). Also, the RCT methodology in a study that generates nonsignificant effects for an intervention may, in fact, obscure positive features and/or outcomes of the intervention which go undetected (Norwich & Koutsouris, 2020). Evidently, this was such a concern that the EER now requires applicants to employ a parallel "process study" that accompanies any RCT (Norwich & Koutsouris, 2020). Such a requirement here would move the field toward mixed-methods research and establish appropriate practices and criteria for conducting and evaluating that research (Hitchcock et al., 2018).

Process evaluations are now employed frequently in medical and public health research to determine the ways in which CSIs lead or do not lead to proposed outcomes (Moore et al., 2015). They typically examine the fidelity of the intervention's implementation, the "reach" of the study (i.e., who the participants were and their context), participants' appraisal of the intervention, and the service paths that led to outcomes. From the field of public health, Limbani et al. (2019) reviewed the process evaluations of seven international RCT studies of hypertension interventions, identifying a variety of procedural and theoretical

approaches. In an example from education science, Siddiqui et al. (2018) conducted process evaluations of two high school literacy interventions. Their findings indicated the interventions to be beneficial for students, not only documented involvement of the teachers as implementers and evaluators but also documented that study protocols may not have always been consistently followed which informed the interpretation of the outcome data.

Design-Based Experimentation and Improvement Science

Design-based experimentation, which originated in engineering, was applied initially in education by Ann Brown (1992), takes process orientation from description to experimentation. It directly involves the practitioner in employing the intervention, gathers information about effects on children and implementers, and allows latitude for adaptation and accommodation to the local context (Burkhardt & Schoenfeld, 2003). From early work by Brown and her colleagues, the design-based approach has evolved into a field now called Improvement Science (Lewis, 2015), which takes practitioners and schools through "improvement cycles" that are designed to fit the intervention to the school in a systematic but data-informed way. Design-based experimentation is often relegated to a development phase of an intervention's validation that will then be "tested" in an RCT efficacy trial. However, it could be employed proactively as an experimental approach to establishing the external validity of a known intervention on a case-by-case basis.

Single-Case Design

In the larger discussions of RCT methodology and evidence-based practices in disciplines outside of special education, SCD is rarely if ever mentioned as an alternative methodology. To its credit, the IES funded What Works Clearinghouse and the two national centers on education and special education research recognize SCD as an experimental methodology that can provide evidence of practice efficacy. However, even within education, meta-analyses exclude SCD studies or assign such studies a lower methodological quality (Aha et al., 2012). High-quality SCDs are experimental in that they address nearly all the threats to internal validity that Campbell and Stanley (1963) initially established and that remain the guiding markers today (Kazdin, 2011). The title, *SCD*, implies to the uninformed that the study includes only one participant, which rarely is true. In fact, SCDs can involve whole classrooms of students, whole schools, or even whole communities. Another criticism of SCD is that statistical analysis is not employed, and it is true that visual inspection is a primary mode of data analysis. However, currently, there are a variety of options for employing statistical analyses for SCD (Shadish et al., 2015). The most common criticisms from group-design methodologists appear to be related to the limited generalizability or external validity of SCD, but as noted previously, this is a major criticism of RCTs also.

Replications

External validity is the Achilles heel of RCTs. Any single RCT study is context bound (Joyce & Cartwright, 2020). So, in most cases, the findings of a study can only apply to those students, teachers, schools, and communities involved in the study (Deaton, 2020). External validity is established through replication. The number and type of replications depend on the population to which one would want to generalize the findings. In special education, to show that a practice is evidence-based requires at least two RCTs, four quasi-experimental group comparison studies, or five SCD studies (Council for Exceptional Children [CEC], 2014). Criteria such as these are strengthened if the replications are conducted by independent research groups (Hume et al., 2021), which CEC unfortunately omitted from their standards for evidence-based practices.

For many years, direct replication studies were not encouraged. They generally were undertaken when a controversial intervention or treatment was in question. For example, facilitated communication is one of the most controversial interventions in special education. Unsubstantiated claims by advocates led to many replication studies, or rather failure-to-replicate studies, for facilitated communication (Holehan & Zane,

2020). It is encouraging that the field is currently placing greater emphasis on direct replications. For example, IES now has a research competition that funds replications of studies that have initial evidence of efficacy. Also, journals are now more open to publishing replication studies than at any time in the past (e.g., the *Journal of Applied Behavior Analysis* has a special section set aside for replications). This is a positive step toward addressing the external validity issue in educational research.

Mixed-Method Research

To this point, concerns about the singular use of RCTs have been voiced, but their relationship to other research methodologies has not been addressed. Recognizing that a single type of methodology may not answer the important questions in education, Greene et al. (1989) charted a course for integrating (i.e., mixing) multiple research methods to provide complementary information from which to draw conclusions. They defined mixed-method research as "those that include at least one quantitative method (designed to collect numbers) and one qualitative method (designed to collect words), where neither type of method is inherently linked to any particular inquiry paradigm" (p. 256). Mixed-method research has elaborated in sophistication and precision in the last three decades (Creswell & Clark, 2017), in part because of the argument that RCTs alone may not provide answers to the critical questions for intervention evaluation research (Fetters & Molina-Azorin, 2020). As an example, in a mixed-method study of a school-based program to enhance self-efficacy and reduce depression, Mackay et al. (2017) employed both an RCT and interventions of student participants. From the quantitative parent-rating measures, they found positive treatment effects but not from students' self-ratings; however, from student interventions, they found convincing evidence of treatment effects, which could be integrated with the quantitative findings. The authors integrated the two sets of information in discussing the relevant effects of the intervention for specific students from specific perspectives and in different contexts. If only the quantitative measures had been employed, important treatment effects would have been missed.

Summary and Conclusion

This article began by indicating the importance of research to identify practices that are effective for specific individuals in specific contexts and delivered by specific individuals. I have described the hesitations that researchers from a variety of disciplines have in adopting RCTs as a gold standard methodology that might not only address those issues but also suggest that RCTs could have an important place in some research endeavors when integrated with other methodological approaches. Other types of studies (e.g., procedural studies which often are mixed in methodology, SCDs, design research), I propose, will also have great value in this research endeavor. In a quote attributed to different individuals but that I first heard from Fixsen et al. (2015), "all organziations and systems are perfectly designed to achieve exactly the results they obtain" (p. 702). As a field, we are following a gold standard methodology that primarily generates effect sizes for educational interventions that are hardly better than standard educational practice (Edovald & Nevill, 2021). Certainly, exceptions exist (e.g., Bradshaw et al., 2010), but they are the exception rather than the rule. A future direction could be to examine the premises on which RCT methodology is based and look for complementary or alternative research methods that may increase the relevance of research outcomes for children and youth with and without disabilities.

Acknowledgment

The author thanks Dr. Sallie Nowell and the two editors for their assistance with the final production of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Samuel L. Odom https://orcid.org/0000-0003-1745-7915

Note

1. There are actually other randomized experimental designs. For the sake of discussion and space constraints, and because it is the primary randomized controlled trial design used in education, only the randomized pretest–posttest control group design (D. T. Campbell & Stanley, 1963) will be reviewed here.

References

Aha, S., Ames, A., & Myers, N. (2012). A review of meta-analyses in education. *Review of Educational Research*, *82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Bédécarrats, F., Guérin, I., & Roubaud, F. (2017). All that glitters is not gold: The political economy of randomized evaluations in development. *Development and Change*, *50*(3), 735–762. https://doi.org/10.1111/dech.12378

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18–20. https://doi.org/10.3102/0013189X031008018

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomized controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, *75*(12), 2299–2306. https://doi.org/10.1016/j.socscimed.2012.08.032

Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). Assessing the gold standard–Lessons from the history of RCTs. *New England Journal of Medicine*, *374*(22), 2175–2181. https://doi.org/10.1056/NEJMms1604593

Boyd, B. A., Hume, K., McBee, M. T., Alessandri, M., Guitierrez, A., Johnson, L., Sperry, L., & Odom, S. L. (2014). Comparative efficacy of LEAP, TEACCH and non-model-specific special education programs for preschoolers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *44*, 366–380. https://doi.org/10.1007/s10803-013-1877-9

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of school-wide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148. https://doi.org/10.1177/1098300709334798

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, *2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Bruhn, M., & McKenzie, D. (2008). *In pursuit of balance: Randomization in practice in development field experiments*. Policy Research Working Paper No. 4752. World Bank. https://openknowledge.worldbank.org/handle/10986/6910

Burkhardt, H., & Schoenfeld, A. (2003). Improving educational research: Toward a more useful, more influential, and better-grounded enterprise. *Educational Researcher*, *32*(4), 3–14. https://doi.org/10.3102/0013189X032009003

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, *321*(7262), 694–696. https://doi.org/10.1136/bmj.321.7262.694

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/S1745255207005029

Cartwright, N. (2019). *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Open Court.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*. Nuffield Trust.

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. *Journal of Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Council for Exceptional Children. (2014). *Council for exceptional children standards for evidence-based practices in special education*.

Creswell, J. W., & Clark, V. L. (2017). *Designing and conducting mixed method research*. SAGE.

Deaton, A. (2020). Randomization in the tropics revisited: A theme and eleven variations. In F. Bédécarrats, I. Guérin, & F. Roubaud (Eds.), *Randomized controlled trials in the field of development: A critical perspective* (pp. 29–46). Oxford University Press.

Edovald, T., & Nevill, C. (2021). Working out what works: The case of the Education Endowment Foundation in England. *ECNU Review of Education*, *4*(1), 46–64. https://doi.org/10.1177/2096531120913039

Education Sciences Reform Act of 2002 (ESRA, Title I of P.L. 107-279).

European Commission. (2016). *Counterfactual impact evaluation*. Joint Research Center. https://ec.europa.eu/jrc/en/research-topic/counterfactual-impact-evaluation

Fenton, W. S. (2000). A programmatic approach to socially complex intervention development. *Journal of Mental Health Policy and Economics*, *3*(2), 113–114. https://.doi.org/10.1002/1099-176x(200006)3:2<113::aid-mhp75>3.0.co;2-h

Fetters, M. D., & Molina-Azorin, J. F. (2020). Utilizing a mixed method approach for conducting interventional research. *Journal of Mixed Method Research*, *14*(2), 131–144. https://doi.org/10.1177/1558689820912856

Fixsen, D., Blase, K., Metz, A., & Van Dyke, M. (2015). Implementation science. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 695–702). Elsevier.

Frey, K. S., Nolen, S. B., Edstrom, L. V., & Hirschstein, M. K. (2005). Effects of a school-based social-emotional competence program: Linking children's goals, attributions, and behavior. *Journal of Applied Developmental Psychology*, *26*(2), 171–200. https://doi.org/10.1016/j.appdev.2004.12.002

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. https://doi.org/10.3102/01623737011003255

Greenhalgh, T. (2000). *How to read a paper: The basics of evidence based medicine*. John Wiley.

Hitchcock, J. H., Johnson, R. B., & Schoonenboom, J. (2018). Idiographic and nomothetic causal inference in special education research and practice: Mixed methods perspectives. *Research in the Schools*, *25*(2), 56–67.

Holehan, K. M., & Zane, T. (2020). *Is there science behind that? Facilitated communication*. Association for Science in Autism Treatment. https://asatonline.org/for-parents/becoming-a-savvy-consumer/is-there-science-behind-that-facilitated-communication/

Hume, K., Boyd, B., McBee, M., Coman, D., Gutierrez, A., Shaw, E., Sperry, L., Alessandri, A., & Odom, S. (2011). Assessing implementation of comprehensive treatment models for young children with ASD: Reliability and validity of two measures. *Research in Autism Spectrum Disorders*, *5*(4), 1430–1440. https://doi.org/10.1016/j.rasd.2011.02.002

Hume, K., Steinbrenner, J. R., Odom, S. L., Morin, K. L., Nowell, S. W., Tomaszewski, B., Szendrey, S., McIntyre, N. S., Yücesoy-Özkan, S., & Savage, M. N. (2021). Evidence-based practices for children, youth, and young adults with autism: Third generation review. *Journal of Autism and Developmental Disorders*. Advanced online publication. https://doi.org/10.1007/s10803-020-04844-2

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, *57*(3), 1045–1082. https://doi.org/10.3102/0002831219866687

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). Oxford University Press.

Kernan, W. N., Viscoli, C. M., Makuch, U. W., Brass, L. M., & Horowitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*, *52*(1), 19–26. https://doi.org/10.1016/s0895-4356(98)00138-3

Kuklick, H., & Kohler, R. E. (1996). Introduction. *Osiris*, *11*, 1–14. http://www.jstor.org/stable/301924

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*(1), 54–61. https://doi.org/10.3102/0013189X15570388

Limbani, F., Groudge, J., Joshi, R., Maar, M. A., Miranda, J. J., Oldenburg, B., Parker, G., Pesantes, M. A., Riddell, M. A., Salam, A., Trieu, K., Thrift, A. G., Van Olmen, J., Vedanthan, R., Webster, R., Yeates, K., & Webster, J., & Global Alliance for Chronic Diseases, Process Evaluation Working Group. (2019). Process evaluation in the field: Global learnings from seven implementation research hypertension projects in low-and middle-income countries. *BMC Public Health*, *19*, Article 153. https://doi.org/10.1186/s12889-019-7261-8

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms (NCSER 2013-3000)*. National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

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

Mackay, B. A., Shochet, I. A., & Orr, J. A. (2017). A pilot randomised controlled trial of a school-based resilience intervention to prevent depressive symptoms for young adolescents with autism spectrum disorder: A mixed methods

analysis. *Journal of Autism and Developmental Disorders*, *47*(11), 3458–3478. https://doi.org/10.1007/s10803-017-3263-5

Marchal, B., Westhorp, A., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G., & Pawson, R. (2013). Realist RCTs of complex interventions: An oxymoron. *Social Science & Medicine*, *94*, 124–128. http://doi.org/10.1016/j.socscimed.2013.06.025

McGaghie, W. C. (2011). Implementation science: Addressing complexity in medical education. *Medical Teacher*, *33*(3), 97–98. https://doi.org/10.3109/0142159X

McKnight, L., & Morgan, M. (2020). A broken paradigm? What education needs to learn from evidence-based medicine. *Journal of Education Policy*, *35*(5), 648–664. https://doi.org/10.1080/02680939.2019.1578902

McPherson, A., Saltmarsh, S., & Tomkins, S. (2020). Reconsidering assent for randomised control trials in education: Ethical and procedural concerns. *British Educational Research Journal*, *46*(4), 728–746. https://doi.org/10.1002/berj.3624

Methods Group of the Campbell Collaboration. (2016). *Methodological expectations of Campbell Collaboration intervention reviews: Conduct standards* (*Campbell Policies* and Guidelines Series No. 3). https://doi.org/10.4073/cpg.2016.3

Mezey, G., Robinson, F., Campbell, R., Gillard, S., Macdonald, G., Meyer, D., Bonell, C., & White, S. (2015). Challenges to undertaking randomised trials with looked after children in social care settings. *Trials*, *16*, Article 206. https://doi.org/10.1186/s13063-015-0708-z

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonnell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation and complex interventions: Medical Research Council guidance. *British Medical Journal*, *350*, Article h1259. https://doi.org/10.1136/bmj.h1258

National Research Council. (2002). *Scientific research in education*. The National Academies Press. https://doi.org/10.17226/10236

Norwich, B., & Koutsouris, G. (2020). Putting RCTs in their place: Implications from an RCT of the integrated group reading approach. *International Journal of Research & Method in Education*, *43*(2), 113–126. https://doi.org/10.1080/1743727X.2019.1626820

Odom, S. L., Duda, M. A., Kucharczyk, S., Cox, A. W., & Stabel, A. (2014). Applying an implementation science framework for adoption of a comprehensive program for high school students with autism spectrum disorders. *Remedial and Special Education*, *35*(2), 123–132. https://doi.org/10.1177%2F0741932513519826

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D. A., & Mittman, B. S. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health and Mental Health Services Research*, *36*, 24–34. https://doi.org/10.1007/s10488-008-0197-4

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal*, *312*(7023), 71–72. https://doi.org/10.1136/bmj.312.7023.71

Sam, A. M., Odom, S. L., Tomaszewski, B., Perkins, Y., & Cox, A. W. (2021). Employing evidence-based practices for children with autism in elementary schools. *Journal of Autism and Developmental Disorders*, *51*, 2308–2323. https://doi.org/10.1007/s10803-020-04706-x

Schneider, M. (2019, January). *Opening plenary: IES director welcome* (Paper presentation). *IES principal investigators meeting*, *Washington, DC*.

Schwarz, A. (2014). *Experiments in practice*. Pickering & Chatto.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research*. Institute of Education Sciences. https://ies.ed.gov/ncser/pubs/2015002/pdf/2015002.pdf

Siddiqui, N., Gorard, S., & See, B. H. (2018). The importance of process evaluation for randomised control trials in education. *Educational Research*, *60*(3), 357–370. https://doi.org/10.1080/00131881.2018.1493349

Steinbrenner, J. D., Odom, S. L., Hall, L. J., & Hume, K. A. (2020). Moving beyond fidelity: Assessing implementation of a comprehensive treatment program for adolescents with autism spectrum disorder. *Exceptional Children*, *86*(2), 137–145. https://doi.org/10.1177/0014402919855321

Styles, B., & Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research: Methodological debates, questions, challenges. *Journal of Educational Research*, *60*(3), 255–264. https://doi.org/10.1080/00131881.2018.1500194

Wahlberg, A., & McGoey, L. (2007). An elusive evidence base: The construction and governance of randomized controlled trials. *BioSocieties*, *2*(1), 1–10. https://doi.org/10.1017/S1745855207005017

Weisburd, D., Petrosino, A., & Fronius, T. (2014). Randomized experiments in criminology and criminal justice. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 49–63). Springer.

Wolff, N. (2000). Using randomized controlled trials to evaluate socially complex services: Problems, challenges, and recommendations. *Journal of Mental Health Policy and Economics*, *3*(2), 97–109. https://doi.org/10.1002/1099-176x(200006)3:2<97::aid-mhp77>3.0.co;2-s

Woolfolk, R. L. (2015). Clinical trials in psychiatry and clinical psychology: Science or product testing? *Acta Psychopathologica*, *1*(2), 12. https://doi.org/10.4172/2469-6676.100012

Author Biography

Samuel L. Odom is a Senior Research Scientist at the Frank Porter Graham Child Development Institute, University of North Carolina at Chapel Hill; Adjunct Professor at San Diego State University; and Senior Lecturer in the Specialpedagogiska institutionen Stockholms universitet. His research, with colleagues, addresses evidence-based practices and school-based programs for children and youth with autism.

Date Received: November 5, 2020
Date of Final Acceptance: March 18, 2021
Editor-in-Charge: Robert H. Horner