Implementation Methods and Measures

Dean L. Fixsen, Melissa K. Van Dyke, Karen A. Blase

Active Implementation Research Network

Please cite as:

Fixsen, D. L., Van Dyke, M. K., & Blase, K. A. (2019). Implementation methods and measures. Chapel Hill, NC: Active Implementation Research Network.
www.activeimplementation.org/resources

Implementation Methods and Measures

Dean L. Fixsen, Melissa K. Van Dyke, Karen A. Blase

Implementation research is known to involve multiple variables operating at multiple levels to influence implementation outcomes and eventual innovation outcomes. Peters, Adam, Alonge, and Agyepong Akua (2013, p. abstract) described the multiple influences at work in implementation research environments:

> Implementation research seeks to understand and work within real world conditions, rather than trying to control for these conditions or to remove their influence as causal effects. What this means is working with populations that will be affected by an intervention, rather than choosing beneficiaries who may not represent the target population of an intervention. Context has a big role in implementation research as well and can include the social, cultural, economic, political, legal, and physical environment, as well as the institutional setting, comprising various stakeholders and their interactions, and the demographic and epidemiological conditions. Implementation research is also especially concerned with the users of the research and not purely the production of knowledge.

Goggin (1986) adds that under these realistic conditions the number of relevant variables quickly outstrips the number of cases and degrees of freedom. Goggin advises researchers to test theories with small N studies to identify functional elements of an overall theory then combine the functional elements in a large N study to evaluate overall and interaction effects among functional elements.

Greenhalgh, Robert, MacFarlane, Bate, and Kyriakidou (2004, p. 615) caution that "Context and 'confounders' lie at the very heart of the diffusion, dissemination, and implementation of complex innovations. They are not extraneous to the object of study; they are an integral part of it. The multiple (and often unpredictable) interactions that arise in particular contexts and settings are precisely what determine the success or failure of a dissemination initiative."

The science of implementation has been slow to develop as researchers learn how to do relevant research in complex environments where multilevel influences are a part of every study and confound the development of independent variables and assessment of dependent variables.

Research Methods

Advancing implementation as a science requires experiments that test predictions derived from theory. For present purposes, the focus is on experimental methods and measures. Sampson (2010, p. 491) states that while a common (mis)understanding links causality to particular experimental methods, science "fundamentally is about principles and procedures for the systematic pursuit of knowledge involving the formulation of a problem, the collection of data through observation or experiment, the possibility of replication, and the formulation and testing of hypotheses."

Commonly used research methods to conduct research regarding implementation include:

- Observation / Participant Observation
- Surveys
- Interviews
- Focus Groups
- Experiments
- Secondary Data Analysis / Archival Study
- Mixed Methods (combination of some of the above)

Mixed methods are commonly encouraged by those doing implementation research in attempts to account for many factors while studying more intensely a few factors (Aarons, Fettes, Sommerfeld, & Palinkas, 2012; Bergman & Beck, 2011; Chatterji, 2006; Palinkas et al., 2011; Palinkas et al., 2015; Peters et al., 2013; Teal, Bergmire, Johnston, & Weiner, 2012).

While randomized control trials (RCTs) are viewed by many as the "gold standard" for experimental research, Handley, Lyles, McCulloch, and Cattamanchi (2018, pp. 6-7) point out that traditional RCTs strongly prioritize internal validity over external validity. However:

> In real-world settings, random allocation of the intervention may not be possible or fully under the control of investigators because of practical, ethical, social, or logistical constraints. For example, when partnering with communities or organizations to deliver a public health intervention, it may not be acceptable that only half of individuals or sites receive an intervention. As well, the timing of intervention rollout may be determined by an external process outside the control of the investigator, such as a mandated policy. Also, when self-selected groups are expected to participate in a program as part of routine care, ethical concerns associated with random assignment would arise, for example, the withholding or delaying of a potentially effective treatment or the provision of a less effective treatment for one group of participants." [Under these circumstances] "quasi-experimental designs (QEDs), which first gained prominence in social science research, are increasingly being employed to fill this need. QEDs test causal hypotheses but, in lieu of fully randomized assignment of the intervention, seek to define a comparison group or time period that reflects the counterfactual (i.e., outcomes if the intervention had not been implemented).

Multiple-baseline design

Quasi-experimental designs such as multiple-baseline designs (MBD) have been used in applied research for many years (Braukmann, Kirigin Ramp, Braukmann, Willner, & Wolf, 1983; Caron & Dozier, 2019; Dancer et al., 1978; Embry et al., 2004; Willner et al., 1977). Full descriptions of the MBD can be accessed in summaries of research methods (Horner et al., 2005; Kazdin, 1982, 2002). The value of the MBD is that it first demonstrates and then replicates a functional relationship between an intervention and its outcomes. Typically, three or more baselines are established with a similar person, group, organization, or system as the participant in each baseline. Three is a recommended minimum number of baselines: one demonstration and two replications. An intervention is introduced to each group in a planned sequence (not all at the same time). Thus, as shown in Figure 1, each baseline serves as a control for the others.

The MBD is the design of choice for establishing a science of implementation. First, if a functional relationship (if this, then that) cannot be demonstrated and replicated with just a few participants then there is no reason to do an elaborate group design. For example, Barrish, Saunders, and Wolf (1969) established a functional relationship between the "good behavior game" and improved classroom behavior of students. The within subject design (one classroom with students and a teacher) was conducted in 58 school days. The demonstration of a functional relationship (if this, then that) subsequently was tested in RCTs (Kellam, Rebok, Ialongo, & Mayer, 1994; Smith, Osgood, Oh, & Caldwell, 2017) with similar results. The short time periods to establish functional relationships is a major benefit in applied settings where implementation research is done with real people in real time.

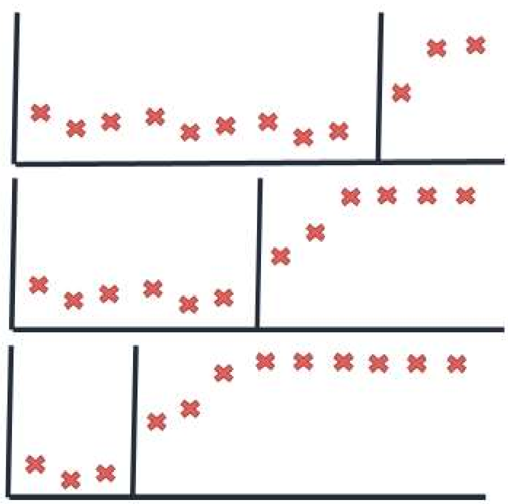

Figure 1. An example of a multiple-baseline design where predictions are tested and replicated in practice.

A second benefit is that the small numbers allow for greater attention to how to produce the independent variable reliably so it is used with fidelity. Rapid adjustments in implementation supports can be made so that high fidelity use of an innovation is available for study (Rahman et al., 2018).

Finally, only a small investment in the research is required before initial results are known. If the use of an innovation in the first baseline produces no change or harmful outcomes, then the experiment can be stopped and none of the other baseline participants will be affected. A MBD can be used to produce reliable information rapidly in environments where "Context and 'confounders' lie at the very heart" of the problems and solutions subject to research.

The MBD produces two forms of comparison data. First, the data show changes in outcomes in each baseline before and after the introduction of an innovation. The first baseline establishes the relationship between an independent variable and dependent variable (if-then) and the next two baselines replicate that pre-post relationship. Second, the staggered introduction of the innovation allows simultaneous comparisons of post intervention scores in baselines with pre intervention scores in the remaining baselines, thus controlling for general trends (e.g. practice effects), time-related events (e.g. funding cuts that affect everyone), and other threats to validity that may affect outcome scores. The MBD eliminates many counterfactual explanations for outcomes associated with interventions (Biglan, Ary, & Wagenaar, 2000; Horner et al., 2005; Jacob, Somers, Zhu, & Bloom, 2016) and resolves the dilemmas encountered when doing research with "too many variables and too few cases" in complex systems (Goggin, 1986).

The MBD is most useful when the dependent variable is relatively stable prior to an intervention (and unlikely to change on its own) and the independent variable is not likely to produce a harmful result (at worst, the intractable behavior will not change). These conditions are encountered routinely in implementation settings where the quality chasm is a common problem.

When an intervention (independent variable) is effective, it is an ethically-preferred design since the participants in each baseline eventually have the opportunity to benefit from the intervention (Doussau & Grady, 2016).

Data analysis using the MBD typically is based on visual inspection of data such as those depicted in Figure 1. To the extent that there is little or no overlap between pre and post intervention scores in each baseline there is little doubt that the intervention produced the predicted outcomes. With little overlap in the distributions of pre scores and post scores, any statistical test would show a significant difference. We strongly suggest displaying MDB and stepped wedge (see below) data in the format shown in Figure 1 with "real time" on the horizontal axis. This provides information about the intervals between data points, overall trends in data points over time in each baseline, overlap in pre-post data points, and the extent to which longer term outcomes in the first baseline are sustained.

Stepped wedge design

In some cases the independent variable is not as powerful as the one illustrated in Figure 1 and more overlap is expected between pre-intervention scores and post-intervention scores. The stepped wedge design employs the MBD logic with statistics used to analyze differences between pre and post innovation scores (Spiegelman, 2016). Mdege (2011) conducted a review of the literature regarding stepped wedge designs. Twenty-five studies (15 study reports and 10 protocols) met the eligibility criteria for inclusion. Twelve of the included studies described the design explicitly as stepped wedge in the title or abstract or within the study protocol description. The remaining 13 studies described their design using one or a combination of phrases such as delayed intervention, delayed treatment, waiting list, phased implementation, phased enrolment, staggered implementation, or crossover. Again, the need for a common language in implementation practice and science is made clear.

In an interesting analysis, Mdege (2011) documented the reasons researchers used a stepped wedge design. Twelve studies cited *methodological reasons* as the motivation for using the stepped wedge design. That is, the design allows for underlying temporal change (n=4), allows multiple comparisons (n=1), addresses existing experimental design deficiencies (n=4), allows generalizability (n=2), allows the use of fewer clinics than would be required in a parallel design (n=1), reduces fear of contamination (n=2), reflects normal practice (n=1), takes advantage of progressive development over time (n=2), and allows the evaluation of the population-level effectiveness of an intervention shown to be effective in an individually randomized trial (n=1). *Logistical reasons* were given in eight studies, which included logistical flexibility (n=2), difficulties in randomizing individuals (n=3), training logistics (n=1), time limitations (n=1), and minimizing researchers' interference with intervention implementation (n=1). Eight studies cited *resource limitations*, including budgetary and human resource constraints. Six studies cited *ethical reasons*, such as a belief that the intervention was effective and individual randomization was ethically questionable. Six studies mentioned *social acceptability reasons*, which were mainly to facilitate study acceptance by participants, the community, and the relevant professionals. *Political acceptability reasons* were cited in four studies, where the intervention under investigation was already an adopted policy but lacked evidence of effectiveness. Most studies were evaluating an intervention during its routine use in practice. For most of the included studies, there was also a belief or empirical evidence suggesting that the intervention would provide more benefits than harm. Denying the intervention to any participant was therefore regarded as unethical or socially/ politically unacceptable.

Mdege (2011) also found that the number of steps (baselines) and participants varied considerably. The number of steps ranged from 2 to 36 steps, with two steps being most common (nine studies). The period between steps also varied considerably from 12 days to 1.5 years. This may reflect the differences in the period from the first use of an innovation to when an observable effect is expected.

The reasons for using a stepped wedge design documented by Mdege (2011) are a good response to the cautions noted in the introduction to this paper (Greenhalgh et al., 2004; Peters et al., 2013). MBD and stepped wedge designs are well suited to cope with research in applied settings where there are too few cases, too many variables, and too little control over multi-level variables that may impact outcomes.

Again, presenting stepped wedge data for visual inspection as shown in Figure 1 should be a requirement. Much more can be learned from seeing the detailed data (e.g., Cummings et al., 2017) than can be learned from a simple statistic (e.g., Trent, Havranek, Ginde, & Haukoos, 2018).

Sampling

Selecting the individuals, groups, organizations, or systems for study in the MBD or stepped wedge design is an important one. The choice is governed by what learning is intended and the circumstances under which a problem is to be solved. Palinkas et al. (2015, pp. 534-536) provide a succinct summary of factors that may affect a sampling decision.

> There exist numerous purposeful sampling designs. Examples include the selection of extreme or deviant (outlier) cases for the purpose of learning from an unusual manifestations of phenomena of interest; the selection of cases with maximum variation for the purpose of documenting unique or diverse variations that have emerged in adapting to different conditions, and to identify important common patterns that cut across variations; and the selection of homogeneous cases for the purpose of reducing variation, simplifying analysis, and facilitating group interviewing.
>
> Embedded in each strategy is the ability to compare and contrast, to identify similarities and differences in the phenomenon of interest. Nevertheless, some of these strategies (e.g., maximum variation sampling, extreme case sampling, intensity sampling, and purposeful random sampling) are used to identify and expand the range of variation or differences, similar to the use of quantitative measures to describe the variability or dispersion of values for a particular variable or variables, while other strategies (e.g., homogeneous sampling, typical case sampling, criterion sampling, and snowball sampling) are used to narrow the range of variation and focus on similarities. The latter are similar to the use of quantitative central tendency measures (e.g., mean, median, and mode). Moreover, certain strategies, like stratified purposeful sampling or opportunistic or emergent sampling, are designed to achieve both goals. As Patton (2002, p. 240) explains, ‘‘the purpose of a stratified purposeful sample is to capture major variations rather than to identify a common core, although the latter may also emerge in the analysis. Each of the strata would constitute a fairly homogeneous sample.’’
>
> In implementation research, quantitative and qualitative methods often play important roles, either simultaneously or sequentially, for the purpose of answering the same question through (a) convergence of results from different sources, (b) answering related questions in a complementary fashion, (c) using one set of methods to expand or explain the results obtained from use of the other set of methods, (d) using one set of methods to develop questionnaires or conceptual models that inform the use of the other set, or (e) using one set of methods to identify the sample for analysis using the other set of methods (Palinkas et al. 2011a).

Implementation Measures

Assessment in practice has been a challenge because of the complexities in human service environments, the novelties encountered in different domains (e.g. education, child welfare, global public health, pharmacy), and the ongoing development of the Active Implementation Frameworks as new research and examined experiences are incorporated into the frameworks.

Pinnock et al. (2017, p. 3) developed useful standards to guide data collection in implementation research. Of the 27 items, #11 states there should be "Defined pre-specified primary and other outcome(s) of the implementation strategy, and how they were assessed. Document any pre-determined targets." And, #12 prompts researchers to specify "Process evaluation objectives and outcomes related to the mechanism(s) through which the strategy is expected to work."

Some of the "mechanism(s) through which the strategy is expected to work" referred to by Pinnock et al. have been proposed in implementation frameworks and measures have been identified. For example, Allen et al. (2017) reviewed the literature related to the "inner setting" of organizations as defined by the Consolidated Framework for Implementation Research (CFIR). Allen et al. (2017) found 83 measures related to the CFIR organization constructs. Consistent with previous findings (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005) only one measure was used in more than one study and the definitions of each construct varied widely across the measures. The two most frequently reported organizational constructs across the studies were "readiness for implementation" (60% of the studies) and "organizational climate" (54% of the studies). These are generic constructs with measures that have been in use for over two decades.

Consequential validity

While the lack of assessment of psychometric properties has been cited as a deficiency by Allen et al. (2017); Lewis et al. (2015), Clinton-McHarg et al. (2016), and others, what is missing from nearly all of the existing implementation-related measures is a test of consequential validity (Shepard, 2016). That is, the information generated by the measure is highly related to using an innovation with fidelity and producing intended outcomes to benefit a population of recipients. Given that implementation practice and science are mission-driven (Fixsen, Blase, Naoom, & Wallace, 2009), consequential validity ("making it happen") is an essential test of any measure.

Galea (2013, p. 1187), working in a health context, stated the problem and the solution clearly:

> A consequentialist approach is centrally concerned with maximizing desired outcomes, and a consequentialist epidemiology would be centrally concerned with improving health outcomes. We would be much more concerned with maximizing the good that can be achieved by our studies and by our approaches than we are by our approaches themselves. A consequentialist epidemiology inducts new trainees not around canonical learning but rather around our goals. Our purpose would be defined around health optimization and disease reduction, with our methods as tools, convenient only insofar as they help us get there. Therefore, our papers would emphasize our outcomes with the intention of identifying how we may improve them.

By thinking of "our methods as tools, convenient only insofar as they help us get there" psychometric properties may be the last thing of concern, not the first (and too often, only) question to be answered. The consequential validity question is "so what?" Once that there is a measure of something it is incumbent on the researcher (the measure developer) to provide data that demonstrates how knowing that information "helps us get there." Once a measure has demonstrated consequential validity then it is worth investing in establishing its psychometric properties to fine tune the measure. It is worth it because it matters.

Measures organized by the Active Implementation Frameworks

Some of the "mechanism(s) through which the strategy is expected to work" referred to by Pinnock et al. have been proposed in the Active Implementation Frameworks. Measures have been developed that are directly related to the factors identified in the Active Implementation Frameworks and the intended outcomes. Other measures have been developed as the Active Implementation Frameworks have been used, revised, and reused in a variety of human service systems. These are known as "action evaluation" measures because they inform action planning and monitor progress toward "making it happen." Action assessments meet the following criteria:

1. They are *relevant* and include items that are indicators of key leverage points for improving practices, organization routines, and system functioning.
2. They are *sensitive* to changes in capacity to perform with scores that increase as capacity is developed and decrease when setbacks occur.
3. They are *consequential* in that the items are important to the respondents and users and scores inform prompt action planning; repeated assessments each year monitor progress as capacity develops.
4. They are *practical* with modest time required to learn how to administer assessments with fidelity to the protocol, and modest time required of staff to respond to rate the items or prepare for an observation visit.

The action evaluation measures listed here are in use by the Active Implementation Research Network and by others in human services.

Table 1. Consequential measures of implementation variables organized by the Active Implementation Frameworks.

Framework	**Concepts**	**Measure**	**Consequential Examples**
Usable Innovation	Innovation Defined Essential Components	Practice profile (Blase, Fixsen, & Van Dyke, 2018)	An innovation that is teachable, learnable, doable, and assessable is more

	Operationalized Components Fidelity Measure	Usability testing (Fraser & Galinsky, 2010)	likely to be used in practice with high fidelity and noticeably improved outcomes for recipients
Implementation Drivers	Competency Drivers Organization Drivers Leadership Drivers	Assessing Drivers Best Practices is a facilitated assessment of implementation supports for practitioners, managers, and leaders in a human service organization. (ADBP; Fixsen, Ward, Blase, et al., 2018)	Higher ADBP scores are associated with high fidelity uses of innovations and sustained outcomes for recipients (Metz et al., 2014; Ogden et al., 2012; Tommeraas & Ogden, 2016)
Organization Drivers	Organizational Culture Organizational Climate Organizational Social Context	Organizational culture, climate, and context surveys that assess organization supports for staff in high demand situations (Glisson, Green, & Williams, 2012; Glisson & Hemmelgarn, 1998; Klein & Sorra, 1996)	Improvements in culture, climate, and context are related to reduced staff turnover, improved morale, and use of effective innovations (Glisson, Dukes, & Green, 2006; Glisson et al., 2010; Glisson et al., 2008; Panzano et al., 2004)
Fidelity	An assessment of the presence and strength of an innovation as it is used in practice by each practitioner	Fidelity assessment is specific to an innovation and assesses the context, content, and	Fidelity is a major contributor to producing benefits to recipients across a variety of innovations

		competence of the use of an innovation through direct observation, record reviews, and asking others (Fixsen et al., 2009; Fixsen et al., 2005)	and service domains (Fixsen, Van Dyke, & Blase, 2019a)
Implementation Stages	Exploration Stage Installation Stage Initial Implementation Stage Full Implementation Stage	A general set of Benchmarks for Stages and innovation-specific Stages of Implementation Completion (Fixsen, Blase, & Van Dyke, 2018; Saldana, Chamberlain, Wang, & Brown, 2012)	Implementation supports can be adjusted to focus on the stage-based requirements related to putting an innovation in place so it can be used with fidelity and sustained (Fixsen et al., 2005; Rahman et al., 2018)
Exploration Stage Installation Stage	ImpleMap interview assesses current implementation strengths in an organization to inform planning the best path toward developing implementation capacity in a specific provider organization	A semi-structured interview process with key respondents to identify current approaches to using innovations in an organization https://www.activeimplementation.org/resources/implemap-exploring-the-implementation-landscape/	ImpleMap results discriminate implementation supports for different innovations in the same organization and for an innovation across organizations https://www.activeimplementation.org/resources/implemap-exploring-the-implementation-landscape/
Installation Stage	Implementation Tracker	An assessment of the presence and strength	The Implementation Quotient is highly

		of an innovation in an organization. Overall fidelity of the use of one or more innovations as the innovation is used in practice by all practitioners in an organization (Fixsen & Blase, 2009)	correlated with overall outcomes for all recipients served by an organization (Fixsen, Blase, & Van Dyke, 2019)
Implementation Teams	Implementation Capacity	Assessments at multiple levels within a system of executive leadership investment, System Alignment, and Commitment to Implementation Team development (Fixsen, Ward, Duda, Horner, & Blase, 2015; Russell et al., 2016; St. Martin, Ward, Harms, Russell, & Fixsen, 2015; Ward et al., 2015)	Implementation capacity is related to the development of linked implementation teams for scaling innovations in complex state systems (Fixsen, Ward, Ryan Jackson, et al., 2018; Ryan Jackson et al., 2018)
Sustainability	Continued use of innovations with fidelity and intended outcomes	The number of those using an innovation as intended divided by the total number that agreed (attempted) to use the innovation (Panzano et al., 2004)	Patterns of sustained use of innovations have been identified and linked to implementation supports and organization factors (Fixsen & Blase, 2018; Massatti,

			Sweeney, Panzano, & Roth, 2008; McIntosh, Mercer, Nese, & Ghemraoui, 2016)
Scaling	Use of innovations with fidelity and intended outcomes for a population of recipients	The number of those receiving an innovation used as intended divided by the total number (the population) that can benefit from the use of the innovation (Fixsen, Blase, & Fixsen, 2017)	High fidelity use of effective innovations and effective implementation supports have produced benefits for global and national populations of intended recipients (Fenner, Henderson, Arita, JeZek, & Ladnyi, 1988; Foege, 2011; Tommeraas & Ogden, 2016; Vernez, Karam, Mariano, & DeMartini, 2006)

Summary

Implementation practice is a complex set of interrelated activities that take place in complex environments where interactions between and among individuals, groups, and organizations are difficult to predict or assess. Researchers are learning to embrace the complex and unknowable aspects of these environments and interactions and learning how to conduct relevant and rigorous knowledge under these conditions.

The multiple-baseline design and stepped wedge design are well suited to implementation problems and solutions. They provide a systematic way to introduce implementation independent variables and study their effects in exactly the environments in which implementation supports are (need to be) provided to promote the use of effective innovations. At the same time, practical action assessments of implementation processes and outcomes have been established and their consequential validity has been demonstrated in practice.

A science of implementation is being strengthened with each study that makes a prediction based on theory and tests that prediction in practice (Fixsen, Van Dyke, & Blase, 2019b).

References

Aarons, G. A., Fettes, D. L., Sommerfeld, D. H., & Palinkas, L. A. (2012). Mixed methods for implementation research: application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child Maltreatment, 17*(1), 67-79. doi:10.1177/1077559511426908

Allen, J. D., Towne, S. D., Maxwell, A. E., DiMartino, L., Leyva, B., Bowen, D. J., . . . Weiner, B. J. (2017). Measures of organizational characteristics associated with adoption and/or implementation of innovations: A systematic review. *BMC Health Services Research, 17*(1), 591. doi:10.1186/s12913-017-2459-x

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2*(2), 119-124. doi:10.1901/jaba.1969.2-119

Bergman, D. A., & Beck, A. (2011). Moving from research to large-scale change in child health care. *Academic Pediatrics, 11*(5), 360-368.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*(1), 31-49. doi:10.1023/a:1010024016308

Blase, K. A., Fixsen, D. L., & Van Dyke, M. (2018). *Developing Usable Innovations*. Retrieved from Chapel Hill, NC: Active Implementation Research Network: www.activeimplementation.org/resources

Braukmann, P. D., Kirigin Ramp, K. A., Braukmann, C. J., Willner, A. G., & Wolf, M. M. (1983). The analysis and training of rationales for child care workers. *Children and Youth Services Review, 5*(2), 177-194.

Caron, E., & Dozier, M. (2019). Effects of Fidelity-Focused Consultation on Clinicians' Implementation: An Exploratory Multiple Baseline Design. *Administration and Policy in Mental Health and Mental Health Services Research*. doi:10.1007/s10488-019-00924-3

Chatterji, M. (2006). Grades of Evidence: Variability in Quality of Findings in Effectiveness Studies of Complex Field Interventions. *American Journal of Evaluation, 28*(3), 239-255.

Clinton-McHarg, T., Yoong, S. L., Tzelepis, F., Regan, T., Fielding, A., Skelton, E., . . . Wolfenden, L. (2016). Psychometric properties of implementation measures for public health and community settings and mapping of constructs against the Consolidated Framework for Implementation Research: a systematic review. *Implementation Science, 11*(1), 148. Retrieved from http://dx.doi.org/10.1186/s13012-016-0512-5. doi:10.1186/s13012-016-0512-5

Cummings, M. J., Goldberg, E., Mwaka, S., Kabajaasi, O., Vittinghoff, E., Cattamanchi, A., . . . Lucian Davis, J. (2017). A complex intervention to improve implementation of World Health Organization guidelines for diagnosis of severe illness in low-income settings: a quasi-experimental study from Uganda. *Implementation Science, 12*(1), 126. doi:10.1186/s13012-017-0654-0

Dancer, D., Braukmann, C. J., Schumaker, J., Kirigin, K. A., Willner, A. G., & Wolf, M. M. (1978). The training and validation of behavior observation and description skills. *Behavior Modification, 2*(1), 113-134. doi:10.1177/014544557821007

Doussau, A., & Grady, C. (2016). Deciphering assumptions about stepped wedge designs: the case of Ebola vaccine research. *Journal of Medical Ethics, 42*(12), 797.

Embry, D. D., Biglan, A., Galloway, D., McDaniel, R., Nunez, N., Dahl, M., . . . Nelson, G. (2004). *Evaluation of reward and reminder visits to reduce tobacco sales to young people: A multiple-baseline across two states*. Retrieved from Tucson, AZ: PAXIS Institute:

Fenner, F., Henderson, D. A., Arita, I., JeZek, Z., & Ladnyi, I. D. (1988). *Smallpox and its eradication*. Geneva, Switzerland: World Health Organization.

Fixsen, D. L., & Blase, K. (2009). *Implementation Quotient*. Retrieved from Chapel Hill, NC: Active Implementation Research Network: https://www.activeimplementation.org/resources/implementation-tracker/

Fixsen, D. L., & Blase, K. A. (2018). The Teaching-Family Model: The First 50 Years. *Perspectives on Behavior Science*. doi:10.1007/s40614-018-0168-3

Fixsen, D. L., Blase, K. A., & Fixsen, A. A. M. (2017). Scaling effective innovations. *Criminology & Public Policy, 16*(2), 487-499. doi:10.1111/1745-9133.12288

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19*(5), 531-540. doi:10.1177/1049731509335549

Fixsen, D. L., Blase, K. A., & Van Dyke, M. K. (2018). *Assessing implementation stages*. Retrieved from Chapel Hill, NC: Active Implementation Research Network: www.activeimplementation.org/resources

Fixsen, D. L., Blase, K. A., & Van Dyke, M. K. (2019). *Implementation practice and science*. Chapel Hill, NC: Active Implementation Research Network.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*: National Implementation Research Network, University of South Florida, www.activeimplementation.org.

Fixsen, D. L., Van Dyke, M. K., & Blase, K. A. (2019a). *Implementation science: Fidelity predictions and outcomes*. Retrieved from Chapel Hill, NC: Active Implementation Research Network: www.activeimplementation.org/resources

Fixsen, D. L., Van Dyke, M. K., & Blase, K. A. (2019b). *Science and implementation*. Retrieved from Chapel Hill, NC: Active Implementation Research Network: www.activeimplementation.org/resources

Fixsen, D. L., Ward, C., Blase, K., Naoom, S., Metz, A., & Louison, L. (2018). *Assessing Drivers Best Practices*. Retrieved from Chapel Hill, NC: Active Implementation Research Network https://www.activeimplementation.org/resources/assessing-drivers-best-practices/

Fixsen, D. L., Ward, C., Ryan Jackson, K., Blase, K., Green, J., Sims, B., . . . Preston, A. (2018). *Implementation and Scaling Evaluation Report: 2013-2017*. Retrieved from National Implementation Research Network, State Implementation and Scaling up of Evidence Based Practices Center, University of North Carolina at Chapel Hill:

Fixsen, D. L., Ward, C. S., Duda, M. A., Horner, R., & Blase, K. A. (2015). *State Capacity Assessment (SCA) for Scaling Up Evidence-based Practices (v. 25.2)*. Retrieved from National Implementation Research Network, State Implementation and Scaling up of Evidence Based Practices Center, University of North Carolina at Chapel Hill:

Foege, W. H. (2011). *House on fire: The fight to eradicate smallpox*. Berkeley and Los Angeles, CA: University of California Press, Ltd.

Fraser, M. W., & Galinsky, M. J. (2010). Steps in Intervention Research: Designing and Developing Social Programs. *Research on Social Work Practice, 20*(5), 459-466. doi:10.1177/1049731509358424

Galea, S. (2013). An Argument for a Consequentialist Epidemiology. *American Journal of Epidemiology, 178*(8), 1185-1191. doi:10.1093/aje/kwt172

Glisson, C., Dukes, D., & Green, P. (2006). The effects of the ARC organizational intervention on caseworker turnover, climate, and culture in children's service systems. *Child Abuse & Neglect, 30*(8), 855-880.

Glisson, C., Green, P., & Williams, N. J. (2012). Assessing the Organizational Social Context (OSC) of child welfare systems: Implications for research and practice. *Child Abuse & Neglect, 36*(9), 621. doi:10.1016/j.chiabu.2012.06.002

Glisson, C., & Hemmelgarn, A. (1998). The effects of organizational climate and interorganizational coordination on the quality and outcomes of children's service systems. *Child Abuse and Neglect, 22*(5), 401-421.

Glisson, C., Schoenwald, S. K., Hemmelgarn, A., Green, P., Dukes, D., Armstrong, K. S., & Chapman, J. E. (2010). Randomized trial of MST and ARC in a two-level evidence-based treatment implementation strategy. *Journal of Consulting & Clinical Psychology, 78*(4), 537-550.

Glisson, C., Schoenwald, S. K., Kelleher, K., Landsverk, J., Hoagwood, K. E., Mayberg, S., & Green, P. (2008). Therapist turnover and new program sustainability in mental health clinics as a function of organizational culture, climate, and service structure. *Adm Policy Ment Health, 35*, 124-133.

Goggin, M. L. (1986). The "Too Few Cases/Too Many Variables" Problem in Implementation Research. *The Western Political Quarterly, 39*(2), 328-347.

Greenhalgh, T., Robert, G., MacFarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: Systematic review and recommendations. *The Milbank Quarterly, 82*(4), 581-629.

Handley, M. A., Lyles, C. R., McCulloch, C., & Cattamanchi, A. (2018). Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research. *Annual Review of Public Health, 39*(1), 5-25. doi:10.1146/annurev-publhealth-040617-014128

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children, 71*(2), 165-179. doi:10.1177/001440290507100203

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions. *Evaluation Review, 40*(3), 167-198. doi:10.1177/0193841X16663414

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kazdin, A. E. (2002). *Research design in clinical psychology*. New York: Allyn & Bacon.

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The Course and Malleability of Aggressive Behavior from Early First Grade into Middle School: Results of a Developmental Epidemiologically-based Preventive Trial. *Journal of Child Psychology and Psychiatry, 35*(2), 259-281. doi:10.1111/j.1469-7610.1994.tb01161.x

Klein, K. J., & Sorra, J. S. (1996). The challenge of innovation implementation. *Academy of Management Review, 21*(4), 1055-1080.

Lewis, C., Fischer, S., Weiner, B., Stanick, C., Kim, M., & Martinez, R. (2015). Outcomes for implementation science: an enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science, 10*(1), 155. doi:10.1186/s13012-015-0342-x

Massatti, R., Sweeney, H., Panzano, P., & Roth, D. (2008). The de-adoption of innovative mental health practices (IMHP): Why organizations choose not to sustain an IMHP. *Administration and Policy in Mental Health and Mental Health Services Research, 35*(1-2), 50-65. doi:10.1007/s10488-007-0141-z

McIntosh, K., Mercer, S. H., Nese, R. N. T., & Ghemraoui, A. (2016). Identifying and Predicting Distinct Patterns of Implementation in a School-Wide Behavior Support Framework. *Prevention Science, 17*(8), 992-1001. doi:10.1007/s11121-016-0700-1

Mdege, N. D. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology, 64*, 936-948.

Metz, A., Bartley, L., Ball, H., Wilson, D., Naoom, S., & Redmond, P. (2014). Active Implementation Frameworks (AIF) for successful service delivery: Catawba County child wellbeing project. *Research on Social Work Practice, 25*(4), 415 - 422. doi:10.1177/1049731514543667

Ogden, T., Bjørnebekk, G., Kjøbli, J., Patras, J., Christiansen, T., Taraldsen, K., & Tollefsen, N. (2012). Measurement of implementation components ten years after a nationwide introduction of empirically supported programs – a pilot study. *Implementation Science, 7*, 49. Retrieved from http://www.implementationscience.com/content/pdf/1748-5908-7-49.pdf.

Palinkas, L. A., Aarons, G. A., Horwitz, S., Chamberlain, P., Hurlburt, M., & Landsverk, J. (2011). Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 44-53. doi:10.1007/s10488-010-0314-z

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(5), 533-544. doi:10.1007/s10488-013-0528-y

Panzano, P. C., Seffrin, B., Chaney-Jones, S., Roth, D., Crane-Ross, D., Massatti, R., & Carstens, C. (2004). The innovation diffusion and adoption research project (IDARP). In D. Roth & W. Lutz (Eds.), *New Research in Mental Health* (Vol. 16, pp. 78-89). Columbus, OH: Ohio Department of Mental Health Office of Program Evaluation and Research.

Peters, D. H., Adam, T., Alonge, O., & Agyepong Akua, I. (2013). Implementation research: what it is and how to do it. *BMJ, 347*. doi:10.1136/bmj.f6753

Pinnock, H., Barwick, M., Carpenter, C. R., Eldridge, S., Grandes, G., Griffiths, C. J., . . . for the StaRI Group. (2017). Standards for Reporting Implementation Studies: (StaRI) Statement. *British Medical Journal, 356*(16795). doi:http://dx.doi.org/10.1136/bmj.i6795

Rahman, M., Ashraf, S., Unicomb, L., Mainuddin, A. K. M., Parvez, S. M., Begum, F., . . . Winch, P. J. (2018). WASH Benefits Bangladesh trial: system for monitoring coverage and quality in an efficacy trial. *Trials, 19*(1), 360. doi:10.1186/s13063-018-2708-2

Russell, C., Ward, C., Harms, A., St. Martin, K., Cusumano, D., Fixsen, D., . . . LeVesseur, C. (2016). *District Capacity Assessment Technical Manual*. Retrieved from National Implementation Research Network, University of North Carolina at Chapel Hill:

Ryan Jackson, K., Fixsen, D., Ward, C., Waldroup, A., Sullivan, V., Poquette, H., & Dodd, K. (2018). *Accomplishing effective and durable change to support improved student outcomes*. Retrieved from National Implementation Research Network, University of North Carolina at Chapel Hill:

Saldana, L., Chamberlain, P., Wang, W., & Brown, H. C. (2012). Predicting program start-up using the stages of implementation measure. *Administration and Policy in Mental Health, 39*(6), 419-425. doi:10.1007/s10488-011-0363-y

Sampson, R. J. (2010). Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology. *Journal of Quantitative Criminology, 26*(4), 489-500. doi:10.1007/s10940-010-9117-3

Shepard, L. A. (2016). Evaluating test validity: reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268-280. doi:10.1080/0969594X.2016.1141168

Smith, E. P., Osgood, D. W., Oh, Y., & Caldwell, L. C. (2017). Promoting Afterschool Quality and Positive Youth Development: Cluster Randomized Trial of the Pax Good Behavior Game. *Prevention Science*. doi:10.1007/s11121-017-0820-2

Spiegelman, D. (2016). Evaluating Public Health Interventions: 2. Stepping Up to Routine Public Health Evaluation With the Stepped Wedge Design. *American Journal of Public Health, 106*(3), 453-457. doi:10.2105/AJPH.2016.303068

St. Martin, K., Ward, C., Harms, A., Russell, C., & Fixsen, D. L. (2015). *Regional capacity assessment (RCA) for scaling up implementation capacity*. Retrieved from National Implementation Research Network, University of North Carolina at Chapel Hill:

Teal, R., Bergmire, D., Johnston, M., & Weiner, B. (2012). Implementing community-based provider participation in research: an empirical study. *Implementation Science, 7*(1), 41.

Tommeraas, T., & Ogden, T. (2016). Is There a Scale-up Penalty? Testing Behavioral Change in the Scaling up of Parent Management Training in Norway. *Administration and Policy in Mental Health and Mental Health Services Research, 44*, 203-216. doi:10.1007/s10488-015-0712-3

Trent, S. A., Havranek, E. P., Ginde, A. A., & Haukoos, J. S. (2018). Effect of Audit and Feedback on Physician Adherence to Clinical Practice Guidelines for Pneumonia and Sepsis. *American Journal of Medical Quality, 0*(0), 1062860618796947. doi:10.1177/1062860618796947

Vernez, G., Karam, R., Mariano, L. T., & DeMartini, C. (2006). *Evaluating comprehensive school reform models at scale: Focus on implementation*. Retrieved from Santa Monica, CA: http://www.rand.org/

Ward, C., St. Martin, K., Horner, R., Duda, M., Ingram-West, K., Tedesco, M., . . . Chaparro, E. (2015). *District Capacity Assessment*. Retrieved from National Implementation Research Network, University of North Carolina at Chapel Hill:

Willner, A. G., Braukmann, C. J., Kirigin, K. A., Fixsen, D. L., Phillips, E. L., & Wolf, M. M. (1977). The training and validation of youth-preferred social behaviors of child-care personnel. *Journal of Applied Behavior Analysis, 10*, 219-230.

