Administration and Policy in Mental Health and Mental Health Services Research
https://doi.org/10.1007/s10488-020-01060-z

COMMENTARY

Three Flavorings for a Soup to Cure what Ails Mental Health Services

C. Hendricks Brown[1]

Abstract
With new tools from artificial intelligence and new perspectives on personalizing interventions, we could revolutionize the way mental health services are delivered and achieve major gains in improving the public's mental health. We examine Dr. Bickman's vision around these technological and paradigm changes that would usher in major scientific, workforce training, and societal cultural changes. We argue that additional efforts in research evaluations in implementation have the potential to scale up and adapt existing interventions and scale them out to diverse populations and service systems. The next stage of this work involves testing the effectiveness of personalized interventions that are preferred by the public and integrating these choices into sustainable service systems. We note cautions on the delivery of these programs as automated algorithmic recommendations are heretofore foreign to humans.

Keywords Mental health services · Artificial intelligence · Machine learning · Personalized interventions · Implementation trials · N-of-1 designs · Scaling up · Scaling out · Scaling down

This festschrift paper has much to offer the readers of this journal, more than I would say almost any other festschrift paper that I have read. Dr. Bickman's extraordinary career is well represented in this paper, and the field is deeply indebted to his insights, guidance, and support. I fully applaud the author on not only the paper but also the contributions to the field that he has made over more than 50 years. Inside the two themes on the use of artificial intelligence (AI) and personized interventions there are many insights regarding factors that have limited progress in mental health services. There are many references to work that I was not aware of, and this can be a great starting place for early career investigators who will no doubt spearhead much of the work that is just now being glimpsed in this paper.

As most behavioral scientists have limited familiarity with AI, this paper is an excellent, insightful, challenging, and for some perhaps a terrifying view into the revolutionary future of improved mental health services through the use of artificial intelligence and personalized interventions. The author has had an exceptional career in addressing mental health, and this festschrift paper is a light showing where the current research paradigms, relying heavily on randomized trials, have failed to make the kind of population improvements in mental health that have happened in many areas of physical health, such as reduction in mortality from cardiovascular disease (Mensah et al. 2017). The wisdom of generations has told us that science progresses slowly, but we must be on a tectonic time scale, based on the little progress we see (Weisz et al. 2019). This paper is not for the timid; despite limitations and a melding of methods argument placed later in the paper, there would be real consequences in what clincians, researchers, practitioners, and leaders are taught and how they function if AI were to catch up to or overtake RCTs in guiding mental health services. As the statistician GEP Box once said, "Whole industries of statistics should be shut down." The same shut down, or at least major retooling, needs would be true here if Bickman's perspective on the undeniable need for effective treatment is heeded.

This commentary interacts with the festschrift paper around its two major themes. One is the dissatisfaction with the current paradigm of using randomized clinical trials to improve services. The second is the value of newer

This article is part of the Festschrift for LeonardBickman Special Issue on The Future of Children's Mental Health Services.

✉ C. Hendricks Brown
hendricks.brown@northwestern.edu

[1] Department of Psychiatry and Behavioral Sciences, Northwestern University, 750 N Lake Shore Dr, 10th Floor, Chicago, IL 60611, USA

Published online: 26 July 2020

approaches relying on AI and personalized interventions to improve mental health. Throughout we use the analogy of these two elements as "flavors" to include in a soup where they merge.

The third "flavor" for curing mental health services noted here involves the fundamental human capacity to be empathic, sometimes considered contradictory to the core rational characteristics so embodied by randomized trials and AI. More on that at the end of this commentary.

Randomized trials have progressed since the early days that A.B. Hill persuaded the medical establishment to use them in clinical work (Doll 1992), and my use of this term includes the most innovative trials involving persons, places, and time (Brown et al. 2008). Their benefits, especially to cancer research, and their deficiencies are well noted in Bickman's paper, including the challenges of "scaling up from a rigorous RCT to a community-based treatment" and "scaling down to the individual client level." It is useful to provide some recent updates that complement these themes of scaling up and scaling down.

First Flavor: Large Scale Implementation Studies for Scaling Up and Scaling Out

For the issue of "scaling up," randomized trials have now been expanded to include randomized implementation trials, where the designs involve random assignment of large numbers of health providers, health systems, or counties to alternative implementation strategies. These designs allow us to evaluate which delivery strategies of evidence-based programs is better, faster, or provides larger reach across a wide range of sites (Brown et al. 2014, 2017; Landsverk et al. 2018). Hybrid trials, which combine both implementation and effectiveness evaluations (Curran et al. 2012), help satisfy the concern that those interventions that are designated as evidence-based may lose effectiveness when delivered at scale (Chambers et al. 2013). Recent methodologic developments for implementation include a framework for measuring critical dimensions when "scaling out" an evidence-based program to different populations or delivery systems (Aarons et al. 2017), a typology of implementation strategies (Powell et al. 2014; Waltz et al. 2015), and a logic model for implementation (Smith et al. 2020). Ethical issues involving conducting trials with demonstrated evidence-based preventive or treatment interventions are fundamentally different, as including an arm of the trial not offering a potentially beneficial intervention is far from equipoise. Instead, randomized rollout designs—which include stepped wedge designs—randomize the time that an organization or system is provided supports to implement. These rollout designs can be fair and acceptable to communities and host organizations regardless of whether they are picked to go first or last (Brown et al. 2009, 2006; Wyman et al. 2015). While implementation science has a great deal of methods development to work out for both randomized and non-randomized studies, this developing field has a major capacity to overcome some of the major health disparities in mental health, substance abuse, and other behavioral problems faced by minorities, the poor, and underserved (McNulty et al. 2019; Mensah et al. 2018).

Among the problems that Bickman notes in service system level advances, systems are typically slow to change and often resist obtaining quality data that could be helpful in implementing or improving system response. A component of this is the lack of learning because there is limited feedback of timely, accurate, and unobtrusive monitoring of system level behavior. This is potentially a highly valuable approach to improving implementation fidelity by using inexpensive, unobtrusive measures and carefully designed feedback based on large amounts of data such as text mining, video, or audio. AI can be a real help in this process. The Imel and colleagues' work (Imel et al. 2015) mentioned in this paper on motivational interviewing provides important proof of concepts for this approach, as are the examples of audio extraction used commercially by call centers. An example using such AI tools in implementation research is given by Gallo and colleagues (Berkel et al. 2019; Gallo et al. 2020). Such monitoring and feedback systems can also be developed at a funder level to support large-scale implementation of federal initiatives (Wang et al. 2016). Furthermore, surveys now exist that show promise in predicting which sites are likely to sustain their evidence-based programs after seed funding has ended and can become a core component of a system-level sustainment monitoring system (Palinkas et al. 2019).

We are just at the cusp of designing rigorous evaluations of implementation strategies that can adapt to local contexts. SMART trials (sequential multiple assignment randomized trial) are playing a stronger role in testing individual level interventions that re-randomize individuals to alternative choices when someone has not responded. A few SMART trials are being developed to test what adaptive implementation strategies work best when the organization fails to respond to a standard (Kilbourne et al. 2014). What has not happened yet is the development of SMART implementation trials that permit a wide array of inflection points where re-randomization can occur.

One type of tool that we would predict is going to be used more in the implementation world is simulation experiments based on models of complex behavior. In determining what strategies would optimally improve health over long periods of time, such as the 10-year plan to End the HIV Epidemic (EHE), there is valuable data from current efficacy or effectiveness trials that have identified

evidence-based interventions (e.g., ART and PrEP), but these trials' timeframes are far too short to predict what combination of strategies would achieve the best reduction over the next 10 years. Agent-based models can incorporate data from such experiments and simulate the future under a nearly infinite combination of implementation strategies. These results can then be turned into decision making tools that address the needs and data from local communities (Landsverk et al. 2018).

It almost goes without saying that the best implementation of a weak intervention will not do much to improve mental health problems. But in my view, we have reasonably successful prevention (Brown et al. 2018; Rasing et al. 2017) and treatment programs (Gibbons et al. 2012) for depression and other mental disorders (Kane et al. 2015), but the huge gap we have is that they are not being delivered to those who would likely benefit or when they do, they are often discontinued. Thus improved implementation of what we know would provide benefit is the first flavor we can add to improve mental health.

Second Flavor: AI Approaches to Scaling Down

For the issue of "scaling down," Bickman points to the progress that has been made in personalized or precision interventions in psychiatry and mental health services and prevention. If only we could predict in advance what interventions work best for which individuals, we would be well on our way to improving population mental health. This has been an elusive goal, going back at least to the 1960s when the leading AI researcher in expert systems and founder of biomedical informatics, Ted Shortliffe, envisioned that every physician would choose an optimal intervention by entering a patient's demographic and medical information, receiving a cross-tabulation of how such patients with these characteristics across the nation faired on different treatment regimens, and then picking the intervention that had the best outcome from this table. That decision system never arrived, and for good statistical, informatics, and cultural reasons. His views, as well as the field's views, have evolved since these early days, as informatics, big data, machine learning and statistics, all evolved in response to the US' fractured physical and mental health systems that interacted with confidentiality and financial concerns (Shortliffe 2005).

In the language of randomized trials, the issue of scaling down is viewed in terms of describing who would benefit or who would be harmed by one intervention compared to another, when to deliver it, and for how long. Since individuals in trials are assigned to one intervention (except for crossover trials, more on this later), we cannot explicitly say what each person's response difference would be. Before concluding that randomized trials do not have much new to offer, it is useful to note some of the advances that statistical methods have made. We do have sophisticated approaches to assess some level of moderation through treatment effect modeling (Howe 2019). How useful these approaches are depends on how large the heterogeneity in treatment effect is and its dimensionality. To characterize the degree of variation that can take place, we actually do have some statistical methods that provide some partial descriptions of heterogeneity of treatment effects in the growth modeling framework. Specifically, under relatively mild assumptions, one can decompose an average treatment effect on growth into latent classes that characterize the proportion and trajectory for individuals who benefit from an intervention compared to control, a portion that do not change, and a portion that is harmed. For example, in an antidepressant versus placebo trial of adults, approximately 42% would improve either with or without the drug, 26% would respond only to the drug but not placebo, another 28% would not improve from either, and just a tiny fraction, about 4% would improve slightly on placebo and remain unchanged under the antidepressant (Muthén and Brown 2009). Unfortunately, such methods are most useful in identifying the proportions in these trajectory classes but not the people, because repeatedly they have failed to replicate accurate prediction of who would most likely be in these different classes. Some predictive ability for who could benefit from antidepressive versus CBT can sometimes be obtained reliably (Siddique et al. 2012). Most trials do look at moderating variables to some extent, but few of today's trials are powered to look at more than a handful of such variables. More relevant to the issue of scaling down, such variable level analyses are not likely to capture the underlying set of characteristics, as would a person-level perspective. To examine more than a tiny fraction of variables or person characteristics that could be looked at, single randomized trials are almost always too small and have insufficient power to explore such person-level variations in impact. Would a synthesis of related but independently carried out randomized trials help answer the question of who might benefit or be harmed? Unfortunately, the standard synthesis method of meta-analysis is not very useful for evaluating moderation findings, as trials vary greatly in how they assess and report moderation findings and consequently are hard to calibrate across studies. However, a synthesis of individual-level trial data, i.e. integrative data analysis, from similarly designed trials with similar interventions and related outcome measures can provide much more power to detect both overall and variation in intervention effects (Brown et al. 2013; Dagne et al. 2016). A synthesis of individual level data can also be used to examine variation in mediational pathways (Huang

et al. 2016; Perrino et al. 2016), a major step towards understanding causality.

Bickman notes that algorithmic approaches that AI uses in machine learning, relying on high dimensional data and high dimensional solutions. When the number of units is large, machine learning can provide valuable tools for classification and decision making unavailable to classic statistical procedures. These algorithms can expand our limited human capacity to form classifications, and they have, as Bickman describes, many successes in health, particularly with long time series of health states recorded by ecological momentary assessments or large diagnostic assessment of images. There is no question that such tools could be useful for other media in mental health besides stationary images, including video of therapy sessions (Inoue et al. 2010) and automatic transcription and fidelity ratings of voices from intervention sessions (Gallo et al. 2015). We look forward to these advances. It is noteworthy that the field of AI has indeed made major advances. Expert systems were so successful in mimicking or transcending what experts could do that the development of expert systems is no longer an active area of AI research. Back in the 1960s it was "proven" that machines would never be able to understand a language as complex as English. This was before the development of algorithms based on enormous volumes of text and voice communications now available through most smart phones and automatic voice transcribers used by, for example You Tube. We are now familiar with such voice recognition algorithms in our everyday lives. At present, there are still some major technological challenges. It is extremely difficult to automatically distinguish speakers in group settings; even common machine learning algorithms mistakenly conclude there are many speakers when only a therapist and a client of the same sex and age are present. Another challenge for machine learning is that data reliability and accuracy on service and therapy assessments are often uncomfortably low, and these (and most other) analytic procedures cannot compensate adequately for such poor data, especially when no supervised learning is possible. Other current limitations for machine learning are described by the author; namely the application of causal inference and theory generation in AI. References that are provided suggest optimism that these perspectives could be integrated as new approaches and applications are developed.

Bickman includes an interesting discussion of preferred treatment rules (PTR) or individualized treatment rules (ITR) that use machine learning to compute one's optimal treatment based on data at hand. There is clear research value in conducting a randomized trial where one arm uses the preferred treatment while the other arms are either the same for everyone or are "yoked" to the best-predicted treatment of a matched individual. His example of a yoked design is an interesting idea. Differences between the two responses would then estimate how big a personalized intervention effect could be. One concern is the sample size needed for such personalized trials. If two interventions, for example, have only modestly different overall success rates from each other and the two treatments each are preferred for half the population, then nearly half the study population in the two fixed arms would be assigned their preferred intervention, whether yoked or not. Thus, many comparisons could end up being the same in the assigned and yoked condition, so sample sizes would definitely need to be much larger than that for standard trials. Furthermore, we as a field have continued to neglect client's preference towards one intervention over another, prioritizing what clinicians want to test rather than what clients feel they would benefit from. Trials that incorporate client preferences are very rare (Marcus et al. 2012) but would need to be included in such PTR trials.

There are important ethical issues in yoked trials where someone is knowingly assigned to an intervention that is likely to be suboptimal for them. I wonder what informed consent issues would be needed to conduct such a trial. I'm not sure I would like to be in a trial where someone told me you have half a chance of receiving the intervention we think is best for you, or receiving another intervention that is chosen because it appears best for someone else you don't even know (your yoked partner) that is definitely not you.

In addition, in trials where the outcome is rare, e.g., vaccinations for infectious diseases that have been extraordinarily effective from a public health standpoint, no single person could differentiate the risk of getting an infectious disease if it was 1/1000 without vaccination and 1/10,000 with vaccination, but that is a huge population effect. Preference would be a dominant issue. Even in less dramatic cases, there is the risk that what is labeled as "personalized mental health" does not differ much from person to person and is therefore more a marketing gimmick than a decision support tool.

Statistical significance is easy to find in large samples with tiny effect sizes, whereas personalized treatments would require very high discrimination to be useful. Personalized medicine and All of Us, with its 1 million subjects, is finding that small differences in treatment effects have no significance about treatment outcomes at the individual level, only at population levels. For example, if a standard treatment has 40% response rate and one that is personalized has 50%, 90% of individuals would not receive any additional benefit.

A different approach to personalized intervention is the use of N-of-1 trials, where individuals learn through crossing treatments over time, what is better for themselves. A major proponent of such trials is Naihua Duan, who discusses this approach in a recent paper (Duan 2017). Duan notes that systematic biases can be limited using a sequence of

intervention changes that are assigned by randomization and while no explicit use of AI is proposed, such tools could be used to pare down a potentially large class of interventions into a small collection that could actually be tested in a N-of-1 trial. In this era where mobile health (mHealth) interventions for mental health proliferate, individuals could identify their preferences, then machine learning tools could identify the best available technology enabled solutions (Li et al. 2019; Mohr et al. 2013), and finally a web-based tool could help an individual design a N-of-1 test that would demonstrate which would work best for them. Because of the frequent need to assess mental health status from such a trial that could change interventions every day, it would be essential to incorporate symptom assessments that have varying items, as repeated administration of the same items every day would introduce fatigue and measurement bias over time. This is a good opportunity to use computerized adaptive testing (CAT), which would select new items from a large item bank and efficiently estimate change in symptoms based on prior responses (Gibbons et al. 2013, 2019, 2016). This resurgence of N-of-1 designs could become powerful tools in the hands of end users, contrasted to the earlier work using these designs in the hands of therapist researchers.

At the close of this second flavor of personalization, I wanted to raise one caution. The use of therapy records is noted as a complement to electronic health records, and could contribute to a fuller biosignature. However, there is an underlying caution that should be noted. It is challenging to provide a more valid and reliable automated decision rule than humans can if the data that in the therapy or treatment record are themselves biased. Take, for example, the development of the Columbia Suicide Severity Rating Scale (C-SSRS), which reviewed clinical records from NIMH funded randomized trials—which probably represented some of the highest quality unstructured records from clinicians. Among those notes was a case where a child slapping oneself was considered a suicidal act. The old adage of garbage in/garbage out can apply if we do not continually improve our data quality; in this case, it involves a difficult task of assessing suicide intentionality among children. Perhaps we are never going to improve our core assessments in certain areas. This recognition of the underlying quality of the measures we use needs to be integrated into our delivery of research and service. For example, inactivity as measured by fitbit and mobile phones for personal sensoring has been used to predict depression relapse. However, they need to account for times when these devices are not being worn or carried, in contrast to times when someone is too depressed to get out of bed.

Third Flavor: An Empathic Mental Health Workforce

Continuing with the title's metaphor, this third flavor for a healthful soup already exists, and so it serves as a stock to support the other two flavors of scaling up and scaling down. But there are possibilities that Bickman's transformation could weaken the base of this soup unless we are deliberate in our training. It strikes me that there could be a downside to integrating these great opportunities of novel ways to scale up and scale down, particularly as they involve machine learning to help decide what prevention or treatment one should get, because it is predicted to be better. Psychology, social work, nursing, public health, and other educational programs of mental health professionals have, of course, successfully trained generations of clinicians, guiding their ability to relate empathically to clients, and screened out those few who fail to learn or use the essential skills that engender trust. As machine algorithms become more used in mental health, there is some potential that the base personal relationships between clinicians and clients could be weakened with damaging and lasting effects. Mohr et al. have pointed to front-ending technology enabled interventions with human coaches as a way to build on principles he identifies as supportive accountability (Mohr et al. 2011).

Suppose this country actually decided, as Bickman suggests, to conduct a research agenda using multiple, large trials to test whether AI informed preferred interventions actually improve outcomes compared to either standardized protocols or treatment as usual with all its variations. Suppose further that machines do outshine our current methods. Could we use these trial results to invest in a paradigm shift on how treatments are assigned, and would people accept this new world?

In this time when science has led to major improvements in physical health, there are still deep suspicions held by a large portion of this country that deny its benefits, even to the extent of putting others at risk for failing to vaccinate their young children against highly contagious diseases such as measles. Various reasons for such anti-science bias have been explored in a book entitled "How Superstition Won and Science Lost (Burnham 1987), one of which was the abandonment by scientists of a role that communicated directly with people."

Some scientific principles are inherently counterintuitive to humans. We, including clinicians and researchers, prefer data from the last case we see or the last paper we have read, rather than assembling data in a coherent way. Today, a minority of patients are willing to take part in partially of fully blinded phase III randomized trials where there is some evidence already of safety and biological or behavioral

response. However, humans have even less experience and comfort with algorithms than we have with the concept of medical experimentation. Algorithms have been around for millennia (e.g., Euclid's algorithm for computing the greatest common denominator), but aside from learning how to divide algorithmically, our society has virtually no direct experience with algorithms. It is a new way of thinking for most, and similar to having therapy clients learn new ways of thinking, clinicians know that lecturing is a poor way to learn.

There clearly are opportunities where machine learning algorithms could, if implemented well, fit well, reduce suffering, and save lives, especially by improving surveillance. Recurrent elevations in suicide risk among those who were hospitalized, for example, are extremely variable (Goldston et al. 2008), and could potentially be detected from a system incorporating automated messaging machine learning that was accurate, cost-effective, low burden, and non-intrusive. Other machine learning applications that a few researchers are now investigating would clearly be unethical. For example, efforts are underway to identify individuals in a population who have a higher probability of future criminal behavior and provide this to police. The use of such tools, even if accurate, could damage any credibility that machine learning could have any useful value in the public's eyes.

An integration of AI is going to take some major cultural adjustments to the current community of practitioners, policy makers, researchers, as well as the public. In my experience working with AI experts, mostly from engineering, cultural differences are quite noticeable and there are many misunderstandings. There needs to be a coming together with respect for different disciplines, where each discipline would meld together to form a transdisciplinary team. While psychology and other disciplines are indeed attracting more people with backgrounds in computer science and engineering than in the past, their fit into the traditional subspecialties is not going to be easy. Engineering is not a field I would ordinarily go to to find people who appreciate the empathy and humility needed to establish trust. Nor are those being trained to deliver therapy or ally with clients and organizations around mental health necessarily very open to a mechanistic view of what to deliver to whom and when. These two distinct cultures need to be mixed into a blended soup. Like dialectic behavioral therapy, it takes work to hold these two different worldviews at one time. However, there are specific stages of such mixing that have been identified using cultural exchange theory (Palinkas 2018; Palinkas et al. 2009) that could be used to overcome misunderstandings and monitor progress in this fusion of perspectives to improve mental health.

The thoughtful vision expressed by Bickman is a recipe for our respective fields, one where the boldness in the flavors truly matches this challenge.

Acknowledgements I would like to thank all my colleagues in the Center for Prevention Implementation Methodology (Ce-PIM) for all the many contributions to this commentary and to their willingness to let their disciplines intermingle and coalesce. I also thank the National Institute on Drug Abuse (NIDA) and the NIH Office of Disease Prevention for their support for Ce-PIM (P30DA027828, Brown PI), NIDA for support for sustainment measurement (R34DA037516, Palinkas PI), as well as the National Institute of Mental Health for support on synthesis across trials (R01MH117598, Brown PI). The material in this paper is the responsibility of the author and does not necessarily reflect the opinions of the funders or my colleagues.

References

Aarons, G. A., Sklar, M., Mustanski, B., Benbow, N., & Brown, C. H. (2017). "Scaling-out" evidence-based interventions to new populations or new health care delivery systems. *Implementation Science, 12*(1), 111. https://doi.org/10.1186/s13012-017-0640-6.

Berkel, C., Gallo, C. G., Sandler, I. N., Mauricio, A. M., Smith, J. D., & Brown, C. H. (2019). Redesigning implementation measurement for monitoring and quality improvement in community delivery settings. *Journal of Primary Prevention, 40*(1), 111–127. https://doi.org/10.1007/s10935-018-00534-z.

Brown, C. H., Brincks, A., Huang, S., Perrino, T., Cruden, G., Pantin, H., et al. (2018). Two-year impact of prevention programs on adolescent depression: An integrative data analysis approach. *Prevention Science, 19*(Supplement 1), 74–94. https://doi.org/10.1007/s11121-016-0737-1.

Brown, C. H., Chamberlain, P., Saldana, L., Padgett, C., Wang, W., & Cruden, G. (2014). Evaluation of two implementation strategies in 51 child county public service systems in two states: Results of a cluster randomized head-to-head implementation trial. *Implementation Science, 9*, 134. https://doi.org/10.1186/s13012-014-0134-8.

Brown, C. H., Curran, G., Palinkas, L. A., Aarons, G. A., Wells, K. B., Jones, L., et al. (2017). An overview of research and evaluation designs for dissemination and implementation. *Annual Review of Public Health, 38*(38), 1–22. https://doi.org/10.1146/annurev-publhealth-031816-044215.

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burhart, G., et al. (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. *Prevention Science, 14*(2), 144–156. https://doi.org/10.1007/s11121-011-0207-8.

Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthen, B., et al. (2009). Adaptive designs for randomized trials in public health. *Annual Review of Public Health, 30*, 1–25. https://doi.org/10.1146/annurev.publhealth.031308.100223.

Brown, C. H., Wang, W., Kellam, S. G., Muthén, B. O., Petras, H., Toyinbo, P., et al. (2008). Methods for testing theory and evaluating impact in randomized field trials: Intent-to-treat analyses for integrating the perspectives of person, place, and time. *Drug and Alcohol Dependence, 95*(Suppl 1), S74–S104. https://doi.org/10.1016/j.drugalcdep.2007.11.013.

Brown, C. H., Wyman, P. A., Guo, J., & Peña, J. (2006). Dynamic wait-listed designs for randomized trials: New designs for

prevention of youth suicide. *Clinical Trials, 3*(3), 259–271. https://doi.org/10.1191/1740774506cn152oa.

Burnham, J. C. (1987). *How superstition won and science lost: Popularizing science and health in the United States*. New Brunswick: Rutgers University Press.

Chambers, D. A., Glasgow, R., & Stange, K. (2013). The dynamic sustainability framework: Addressing the paradox of sustainment amid ongoing change. *Implementation Science, 8*, 117. https://doi.org/10.1186/1748-5908-8-117.

Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M., & Stetler, C. (2012). Effectiveness-implementation hybrid designs: Combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical Care, 50*(3), 217–226. https://doi.org/10.1097/MLR.0b013e3182408812.

Dagne, G., Brown, C. H., Howe, G., Kellam, S., & Liu, L. (2016). Testing moderation in network meta-analysis with individual participant data. *Statistics in Medicine, 35*(15), 2485–2502. https://doi.org/10.1002/sim.6883.

Doll, R. (1992). Sir Austin Bradford Hill and the progress of medical science. *BMJ British Medical Journal, 305*(6868), 1521.

Duan, N. (2017). Personalized biostatistics, small data, and N-of-1 trials. In *Prevention Science and Methodology Group Virtual Grand Rounds*.

Gallo, C. G., Berkel, C., Mauricio, A. M., Sandler, I. N., Villamar, J., Mehrotra, S., & Brown, C. H. (2020). Implementation methodology from a systems-level perspective. *Manuscript submitted for publication*.

Gallo, C. G., Pantin, H., Villamar, J., Prado, G. J., Tapia, M., Ogihara, M., et al. (2015). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the familias unidas preventive intervention. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(5), 574–585. https://doi.org/10.1007/s10488-014-0538-4.

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., et al. (2013). The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): A screening tool for depression. *Journal of Clinical Psychiatry, 74*(7), 669–674. https://doi.org/10.4088/JCP.12m08338.

Gibbons, R. D., Hur, K., Brown, C. H., Davis, J. M., & Mann, J. J. (2012). Benefits from antidepressants: Synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Archives of General Psychiatry, 69*(6), 572–579. https://doi.org/10.1001/archgenpsychiatry.2011.2044.

Gibbons, R. D., Smith, J. D., Brown, C. H., Sajdak, M., Tapia, N. J., Kulik, A., et al. (2019). Improving the evaluation of adult mental disorders in the criminal justice system with computerized adaptive testing. *Psychiatric Services, 70*(11), appips201900038. https://doi.org/10.1176/appi.ps.201900038.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12*, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634.

Goldston, D. B., Molock, S. D., Whitbeck, L. B., Murakami, J. L., Zayas, L. H., & Hall, G. C. (2008). Cultural considerations in adolescent suicide prevention and psychosocial treatment. *American Psychologist, 63*(1), 14–31. https://doi.org/10.1037/0003-066X.63.1.14.

Howe, G. W. (2019). Preventive effect heterogeneity: Causal inference in personalized prevention. *Prevention Science, 20*(1), 21–29.

Huang, S., MacKinnon, D. P., Perrino, T., Gallo, C. G., Cruden, G., & Brown, C. H. (2016). A statistical method for synthesizing mediation analyses using the product of coefficient approach across multiple trials. *Statistical Methods & Applications, 25*(4), 565–579. https://doi.org/10.1007/s10260-016-0354-y.

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. *Psychotherapy, 52*(1), 19. https://doi.org/10.1037/a0036841.

Inoue, M., Ogihara, M., Hanada, R., & Furuyama, N. (2010). *Utility of gestural cues in indexing semantic miscommunication.* Paper presented at the 2010 5th International Conference on Future Information Technology.

Kane, J. M., Robinson, D. G., Schooler, N. R., Mueser, K. T., Penn, D. L., Rosenheck, R. A., et al. (2015). Comprehensive versus usual community care for first-episode psychosis: 2-year outcomes from the NIMH RAISE early treatment program. *American Journal of Psychiatry, 173*(4), 362–372. https://doi.org/10.1176/appi.ajp.2015.15050632.

Kilbourne, A. M., Almirall, D., Eisenberg, D., Waxmonsky, J., Goodrich, D. E., Fortney, J. C., et al. (2014). Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): Cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. *Implementation Science*. https://doi.org/10.1186/s13012-13014-10132-x.

Landsverk, J., Brown, C. H., Smith, J. D., Chamberlain, P., Curran, G. M., Palinkas, L., et al. (2018). Design and analysis in dissemination and implementation research. In R. C. Brownson, G. A. Colditz, & E. K. Proctor (Eds.), *Dissemination and implementation research in health: Translating science to practice* (2nd ed., pp. 201–228). New York: Oxford University Press.

Li, D. H., Brown, C. H., Gallo, C., Morgan, E., Sullivan, P. S., Young, S. D., et al. (2019). Design considerations for implementing eHealth behavioral interventions for HIV prevention in evolving sociotechnical landscapes. *Current HIV/AIDS Reports, 16*(4), 335–348. https://doi.org/10.1007/s11904-019-00455-4.

Marcus, S. M., Stuart, E. A., Wang, P., Shadish, W. R., & Steiner, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychological Methods, 17*(2), 244.

McNulty, M., Smith, J. D., Villamar, J., Burnett-Zeigler, I., Vermeer, W., Benbow, N., et al. (2019). Implementation research methodologies for achieving scientific equity and health equity. *Ethnicity and Disease, 29*(Suppl 1), 83–92. https://doi.org/10.18865/ed.29.S1.83.

Mensah, G. A., Cooper, R. S., Siega-Riz, A. M., Cooper, L. A., Smith, J. D., Brown, C. H., et al. (2018). Reducing cardiovascular disparities through community-engaged implementation research: A National Heart, Lung, and Blood Institute Workshop Report. *Circulation Research, 122*(2), 213–230. https://doi.org/10.1161/CIRCRESAHA.117.312243.

Mensah, G. A., Wei, G. S., Sorlie, P. D., Fine, L. J., Rosenberg, Y., Kaufmann, P. G., et al. (2017). Decline in cardiovascular mortality: Possible causes and implications. *Circulation Research, 120*(2), 366–380.

Mohr, D. C., Cheung, K., Schueller, S. M., Brown, C. H., & Duan, N. (2013). Continuous evaluation of evolving behavioral intervention technologies. *American Journal of Preventive Medicine, 45*(4), 517–523. https://doi.org/10.1016/j.amepre.2013.06.006.

Mohr, D. C., Cuijpers, P., & Lehman, K. (2011). Supportive accountability: A model for providing human support to enhance adherence to eHealth interventions. *Journal of Medical Internet Research, 13*(1), e30. https://doi.org/10.2196/jmir.1602.

Muthén, B. O., & Brown, C. H. (2009). Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. *Statistics in Medicine, 28*(27), 3363–3395. https://doi.org/10.1002/sim.3721.

Palinkas, L. A. (2018). *Achieving implementation and exchange: A science of delivering evidence-based practices to at-risk youth In Department of psychiatry ground rounds presentation*. New York: Columbia University.

Palinkas, L. A., Aarons, G. A., Chorpita, B. F., Hoagwood, K., Landswerk, J., & Weisz, J. R. (2009). Cultural exchange and the implementation of evidence-based practice: Two case studies. *Research on Social Work Practice, 19*(5), 602–612. https://doi.org/10.1177/1049731509335529.

Palinkas, L. A., Spear, S., Mendon, S., Villamar, J. A., Reynolds, C., Green, C. D., et al. (2019). Conceptualizing and measuring sustainability of prevention programs and initiatives. *Translational and Behavioral Medicine, 10*(1), 136–145. https://doi.org/10.1093/tbm/ibz170.

Perrino, T., Brincks, A., Howe, G., Brown, C. H., Prado, G., & Pantin, H. (2016). Reducing internalizing symptoms among high-risk, hispanic adolescents: Mediators of a preventive family intervention. *Prevention Science, 17*(5), 595–605. https://doi.org/10.1007/s11121-016-0655-2.

Powell, B. J., Proctor, E. K., & Glass, J. E. (2014). A systematic review of strategies for implementing empirically supported mental health interventions. *Research on Social Work Practice, 24*(2), 192–212. https://doi.org/10.1177/1049731513505778.

Rasing, S., Creemers, D. H., Janssens, J. M., & Scholte, R. H. (2017). Depression and anxiety prevention based on cognitive behavioral therapy for at-risk adolescents: A meta-analytic review. *Frontiers in Psychology, 8*, 1066.

Shortliffe, E. H. (2005). Strategic action in health information technology: Why the obvious has taken so long. *Health Affairs, 24*(5), 1222–1233.

Siddique, J., Chung, J. Y., Brown, C. H., & Miranda, J. (2012). Comparative effectiveness of medication versus cognitive behavioral therapy in a randomized controlled trial of low-income young minority women with depression. *Journal of Consulting and Clinical Psychology, 80*(6), 995–1006. https://doi.org/10.1037/a0030452.

Smith, J. D., Li, D., & Rafferty, M. R. (2020). The implementation research logic model: A method for planning, executing, reporting, and synthesizing implementation projects. *medRxiv*. https://doi.org/10.1101/2020.04.05.20054379.

Waltz, T. J., Powell, B. J., Matthieu, M. M., Damschroder, L. J., Chinman, M. J., Smith, J. L., et al. (2015). Use of concept mapping to characterize relationships among implementation strategies and assess their feasibility and importance: Results from the Expert Recommendations for Implementing Change (ERIC) study. *Implementation Science, 10*(1), 109.

Wang, D., Ogihara, M., Gallo, C. G., Villamar, J., Smith, J. D., Vermeer, W., et al. (2016). Automatic classification of communication logs into implementation stages via text analysis. *Implementation Science, 11*(1), 119. https://doi.org/10.1186/s13012-016-0483-6.

Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., et al. (2019). Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, attention-deficit/hyperactivity disorder, and conduct problems. *Perspectives on Psychological Science, 14*(2), 216–237.

Wyman, P. A., Henry, D., Knoblauch, S., & Brown, C. H. (2015). Designs for testing group-based interventions with limited numbers of social units: The dynamic wait-listed and regression point displacement designs. *Prevention Science, 16*(7), 956–966. https://doi.org/10.1007/s11121-014-0535-6.