Theoretical/Methodological/Review Article

Clinical Psychological Science
2019, Vol. 7(6) 1222–1243

DOI: 10.1177/2167702619858071
www.psychologicalscience.org/CPS

Introducing the Leapfrog Design: A Simple Bayesian Adaptive Rolling Trial Design for Accelerated Treatment Development and Optimization

Simon E. Blackwell[1], Marcella L. Woud[1], Jürgen Margraf[1], and Felix D. Schönbrodt[2]

[1]Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, and [2]Department of Psychology, Ludwig-Maximilians-Universität München

Abstract

The application of basic science research to the development and optimization of psychological treatments holds great potential. However, this process of clinical translation is challenging and time-consuming, and the standard route by which it proceeds is inefficient. Adaptive rolling designs, which originated within cancer treatment research, provide an alternative methodology with potential to accelerate development and optimization of psychological treatments. In such designs, multiple treatment options are tested simultaneously, with sequential Bayesian analyses used to remove poorly performing arms. Further, new treatment arms informed by the latest research findings can be introduced into the existing infrastructure as the trial progresses. These features dramatically reduce the sample sizes needed and offer a means for more rapid and efficient clinical translation. This article outlines the utility of such designs to clinical psychological science, focusing on a new variant termed the *leapfrog* design, and discusses their potential uses to accelerate clinical translation.

Keywords

clinical translation, adaptive trial design, sequential analyses, Bayes factor

Received 8/20/18; Revision accepted 4/22/19

The possibility of developing new treatment approaches in mental health by capitalizing on our knowledge of the basic underlying mechanisms holds much promise. However, this translational process is challenging, inefficient, and slow. The present article considers some of the problems of the translational process as it usually proceeds and outlines an alternative research process involving *adaptive rolling trial* designs as one additional method to promote more efficient clinical translation. A specific variant termed the *leapfrog* design, which has been designed to be particularly useful for psychological treatment development and optimization, is introduced. Via a hypothetical example of a leapfrog trial and discussion of how the basic design could be adapted to different stages of the translational process, this article aims to introduce a research method that could greatly increase the efficiency of clinical translational research.

The Promise and Challenge of Clinical Translation

Clinical translation refers to the process by which treatment development or implementation is informed by basic science research. For example, within the domain of psychological treatment development, experimental research may elucidate a particular cognitive process as playing a key role in maintenance of a disorder, and this may lead to the generation of new or refinement of existing psychological interventions (Holmes et al., 2018). The promise of clinical translation is in the potential to develop

Corresponding Author:
Simon E. Blackwell, Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Bochumer Fenster, Massenbergstraße 9-13, 44787 Bochum, Germany
E-mail: simon.blackwell@rub.de

more effective or scalable treatments, which in the domain of mental health is particularly needed: Mental disorders pose a global health challenge, with depression alone estimated to be a leading cause of disability worldwide (World Health Organization, 2017). The scale of the problem is such that simply scaling up or extending delivery of established treatments such as face-to-face psychological therapies would never be sufficient to meet the need (Kazdin & Blase, 2011). Rather, new treatment approaches are needed that are more efficient, easily disseminable and scalable, and cost-effective. There has already been progress made in these respects, for example via adaptation of psychological therapies such as cognitive behavior therapy (CBT) for Internet delivery (Andersson, Carlbring, & Lindefors, 2016; Andersson & Titov, 2014) or the task-shifting approach of training lay people to deliver interventions, for example in developing countries (e.g., Elbert, Wilker, Schauer, & Neuner, 2017; Patel, Chowdhary, Rahman, & Verdeli, 2011). Within the drive to develop and improve treatments, a particular attraction of translational research is that understanding the underlying mechanisms of a disorder or the recovery process may allow the development of more focused interventions that target these processes directly; such a targeted approach opens up the possibility of focused and efficient minimal interventions that may be less expensive and easier to disseminate (Holmes, Craske, & Graybiel, 2014).

Although the possibility of harnessing basic science findings to develop novel treatment approaches holds great potential, this clinical translational process of bridging theory, basic science, and clinical application poses a major challenge—one that is recognized across all fields of health research (e.g., Woolf, 2008). Efforts to derive a new treatment from cognitive science generally follow a linear process as outlined in, for example, guidance from the United Kingdom Medical Research Council (2000), in which experimental studies are followed by studies of increasing size and clinical application until the intervention is either shown to work or falls at one of the hurdles and is discarded or taken back a step in the investigational process. Although there are success stories emerging from such a translational route, it also has a number of problems (see Fig. 1).

First, it is slow and inefficient: Moving from a basic science idea, through experimental work, first proof-of-principle studies, early-phase efficacy trials, and finally definitive randomized controlled trials (RCTs) is a long, time-consuming, resource-intensive, and expensive process. As a reference point, for drug discovery in mental health, it has been estimated that it takes an average of 13 years to develop a new medication, and often even longer (Joyce, 2014; Rutherford & Roose, 2013). For psychological interventions, in which mechanisms may be particularly complex and difficult to measure and the exact parameters by which they may be robustly targeted can be elusive, it would be unsurprising if this translational process took even longer. Further, treatment optimization becomes more difficult over time as the potential gains over existing approaches become smaller.

Second, the standard translational process carries a high risk of false negatives in that potentially valuable interventions may be discarded on the basis of prematurely conducted RCTs carried out before the optimal parameters or mechanisms have been established. For a number of reasons, researchers may sometimes leap to full-scale RCTs before sufficient preparatory work has been carried out. Particularly in the case of a highly focused intervention targeting a specific process, small changes in task parameters or instructions can easily cause the intervention to miss the target; disappointing results from poorly operationalized experimental studies or premature RCTs can lead erroneously to the conclusion that the whole approach is flawed or ineffective rather than simply being a suboptimal operationalization.

Third, alongside this risk of false negatives, there is great potential for false positives in the early stages of a research line in terms of approaches that look promising along the translational road but at a later phase turn out to be not clinically feasible or for which the mechanism does not translate in the real world. Particularly in cases in which preclinical work is characterized by high risks of bias and nonrepresentative study populations or intervention settings, it may be that a huge amount of time and resources is expended before the lack of viability of a potential treatment is realized.

Although the standard translational framework can be useful for conceptualizing how basic science findings can lead to new intervention development and guide programs of research, the problems highlighted in this section indicate the need to consider how we conduct clinical translational research and how this could be improved.

Initiatives to Improve Translational Psychological Science

In the past few years, there have been an increasing number of efforts and initiatives to improve many aspects of clinical psychological research, including the translational process. For example, one strand of this initiative has focused on improving the methodological quality of studies across the span of translational research, for example, via preregistration, publication of data and materials, emphasis on the importance of independent replication, and improved understanding of the impact of sample size on power and precision (Tackett et al., 2017); this focus in turn has built on initiatives in the broader literature (e.g., Munafò et al., 2017). A focus on mechanisms, from molecular to

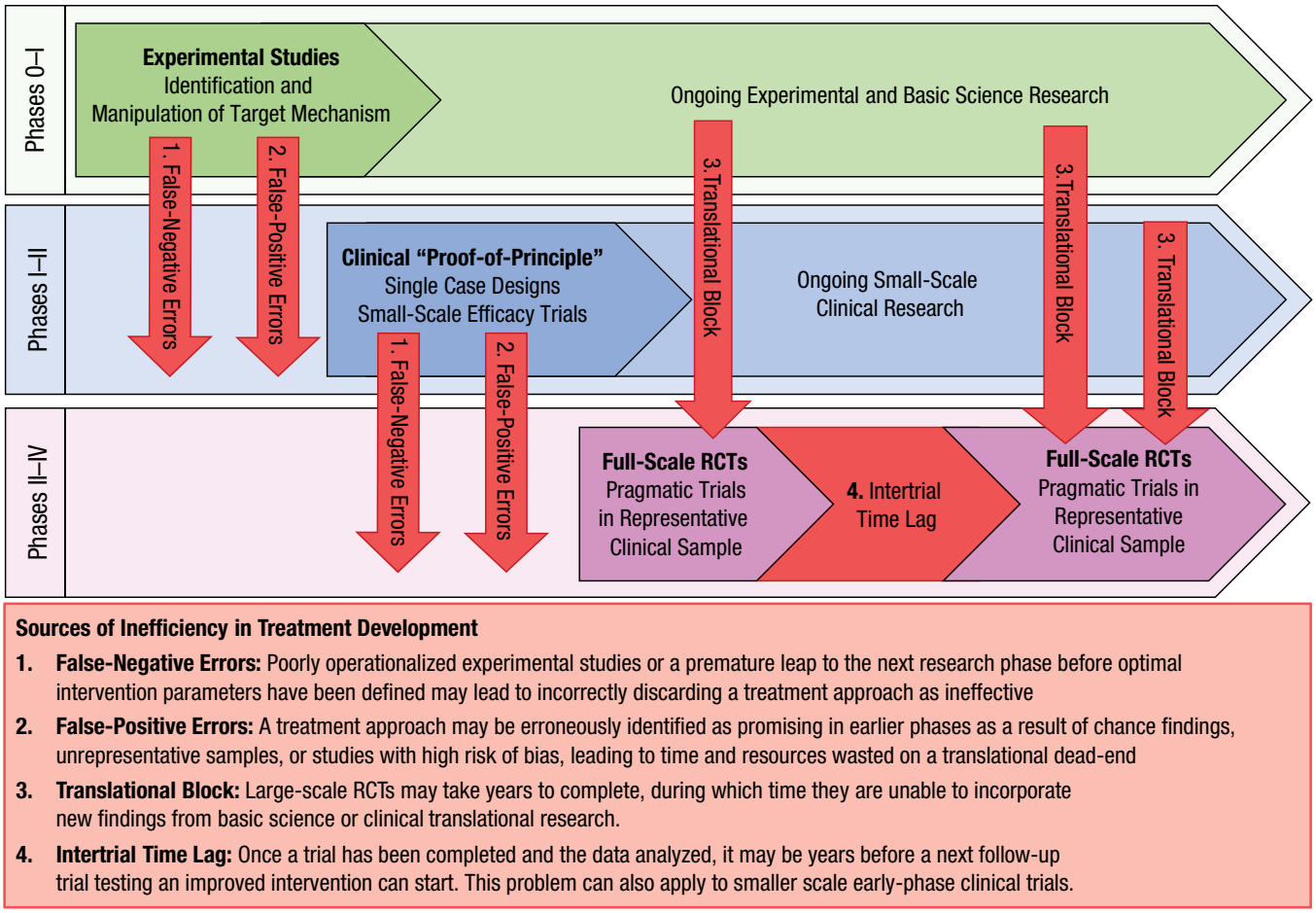

Fig. 1. Sources of inefficiency in treatment development arising from standard translational methodology. Phases 0 to IV refer to those identified in the United Kingdom Medical Research Council (2000) framework. For additional sources of inefficiency related to research waste, arising from factors such as poor research design and publication bias, see Macleod et al. (2014).

societal levels of explanation, alongside improvements in how to conceptualize and measure them has also been advocated (e.g., Holmes et al., 2018; Margraf & Schneider, 2016). Several different approaches to the standard method of two (or three, or more) parallel-arm efficacy trials have been proposed, such as complex factorial designs to assess the impact of different components of a psychological intervention (Watkins et al., 2016). Developing an array of such methodologies that are more suitable for addressing different specific kinds of research questions will greatly enhance treatment development, and it may be helpful to look to other areas of health research for inspiration.

Adaptive Rolling Trial Designs for Psychological Treatment Development

The challenges of clinical translation and treatment development are by no means unique to psychology, and turning to other areas of health research can help provide ideas for methodological innovation. One alternative to the "classical" approach to treatment development, for example as in the Medical Research Council guidelines mentioned above, is inspired by novel designs that have been developed for cancer treatment research. In this field, the problems with standard RCTs in treatment development have long been recognized, spurring the development of alternative methodology (e.g., Hills & Burnett, 2011; Hobbs, Chen, & Lee, 2018; Wason & Trippa, 2014).

In a situation in which a new potential treatment has passed through early-phase testing and is deemed ready for an efficacy RCT against a current established treatment, simply setting up and starting a new RCT takes a huge amount of time and resources in terms of documentation, ethical and other regulatory approvals, assembling and training personnel, and so on. Conducting the trial and collecting outcome data may take several years, during which time other new potential treatments may appear promising or other information

may come to light around mechanisms that may suggest other optimal dosing schedules. However, none of these can be investigated (at least by the research group running the trial) until the current trial has finished and the data have been analyzed (the translational block in Fig. 1), and then there is likely to be a substantial time gap before the new trial is ready to run (the intertrial time lag in Fig. 1). Further, both the data arising from the first trial and that arising from other research published in the meantime may lead to a bewildering array of potential possibilities that could be tested in a follow-up trial. The standard approach to testing treatment possibilities is not only time-inefficient (it has been estimated that testing all current candidate cancer treatments via the standard route would take 90 years; Hobbs et al., 2018) but also wasteful in terms of resources and participant time. For example, if most new treatments turn out in fact not to be superior to the current approach (which from a conservative perspective is perhaps the most likely scenario), many participants are "wasted" testing these new candidate treatments in trials that have to run to completion.

The newer trial designs developed within cancer research can be classed broadly as adaptive rolling trial designs, in which several candidate treatments are tested simultaneously, with interim analyses used to drop treatment arms that appear ineffective; at the same time, novel treatment arms can be added in as new basic science discoveries arise or preliminary clinical work is completed. The idea is to speed up treatment development by feeding new treatment possibilities into an existing trial infrastructure as they become available (e.g., Hobbs et al., 2018; Wason & Trippa, 2014). This reduces the inefficiencies resulting from the translational block and intertrial time lag noted in Figure 1. Further, the ability to rapidly test out a new treatment possibility emerging from basic science research in a real-world application, by capitalizing on an existing trial infrastructure, reduces the potential inefficiencies denoted as false-positive errors in Figure 1. Finally, the ability to drop underperforming treatment arms before completing recruitment can dramatically reduce the sample sizes needed (e.g., by 40% when testing five possible treatments; Hobbs et al., 2018). This practice in turn facilitates testing of several potential variants of a new candidate treatment in initial clinical trials, reducing the potential for false negatives that may arise if only one operationalization of a new treatment can be tested (as noted in Fig. 1). Variants of adaptive rolling designs include those involving repeated null-hypothesis statistical testing (NHST; e.g., Hills & Burnett, 2011), sequential Bayesian analyses (e.g., Hobbs et al., 2018), or adapting the randomization weights such that poorly performing arms are "starved" of participants and drop out (e.g., Wason & Trippa, 2014). The basic ideas of these designs are, of course, not limited to cancer in their potential utility but could be useful across a broad range of research—including that concerning psychological treatment development.

This article aims to illustrate the potential utility of such adaptive rolling trial designs for psychological treatment development and optimization. For the purposes of illustration, it focuses on one simple variant, here termed a *leapfrog* design. Several different ways in which this design could be implemented will be outlined, and a hypothetical example relating to the development of an Internet-delivered cognitive training intervention will be presented. The focus will be on the general principles and how these may be operationalized in concrete terms; readers interested in more detail of the statistical approaches and underlying assumptions or the broader spectrum of adaptive designs available are referred elsewhere (e.g., Hobbs et al., 2018; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).

Using a Leapfrog Design to Develop and Optimize Psychological Treatments

The leapfrog design outlined below is a variant of the platform design (Hobbs et al., 2018), which we have adapted in two main ways for psychological-treatment development and optimization. These adaptations reflect differences between cancer and psychological-treatment research and are intended to make the design and analysis simple to plan and implement by researchers currently engaged in psychological translational research. First, instead of binary outcomes (e.g., tumor presence/absence), the focus is on continuous outcomes (e.g., a score on a questionnaire/clinician rating) using a simple-to-implement analytic method (although the analytic method described could equally be applied to categorical outcomes such as recovered/not recovered if preferred). Second, the design is intended for continuous treatment development and optimization rather than just finding one (or more) treatments better than a current accepted "gold standard." This goal is achieved by building in a mechanism to allow replacement of the comparison arm (the arm against which all other arms are compared) by new treatment arms that achieve a prespecified threshold for demonstrating superiority while the trial is ongoing. The design is here termed a *leapfrog* design in reference to this second design feature in that trial arms can leapfrog (as in the children's game) each other to become the comparison arm or current best. In the following sections, the general principles of the design are described, and an example is then presented to illustrate one potential application.

Design overview

At the start of a leapfrog trial, participants are randomized to two or more trial arms that would each include a different variant of the intervention to be tested. One arm would include the standard or current best version of the intervention, and all comparisons would be made with this arm: The aim is to identify whether new variants have an advantage over the standard. As the trial progresses, after a certain minimum number of participants have been enrolled (here termed N_{min}), a process of continuous monitoring is initiated in which Bayesian modeling is used to calculate the likelihood that a new arm is superior to the standard arm according to a preset criterion (using the method proposed by Schönbrodt et al., 2017, which is based on Bayes factors, BF; see the next section for details).

Using this Bayesian approach allows repeated analyses—even after each participant—without compromising the statistical inference in the way that repeated *p* value–based NHST would (Armitage, McPherson, & Rowe, 1969; Hobbs et al., 2018; Rouder, 2014; Schönbrodt et al., 2017; however, see our limitations section for further discussion). Rather, as data accumulate, the confidence with which one can conclude that the new intervention variant is or is not likely to be superior to the standard version increases. Trial arms in which the new intervention variant does not appear promising, reaching a prespecified threshold (implemented as BF) for sufficient evidence of nonsuperiority (here termed BF_{fail}), are stopped early, meaning that participants are no longer diverted into testing this variant. If a trial arm reaches a predetermined maximum enrollment (here termed N_{max}) without providing sufficient evidence of superiority (reaching a prespecified BF threshold, termed here $BF_{success}$), the arm is also dropped. If a new intervention variant appears superior to the standard arm, with the Bayes factor for comparison with the standard arm reaching a preset criterion $BF_{success}$, the original standard arm is dropped, and the new variant becomes the new standard arm against which new intervention variants are evaluated; thus, any new intervention variant should always be better than the current best. New arms can be added on an ongoing basis, allowing new insights from theory or research to be rapidly tested in this applied context.

Such a trial design could complement and run alongside standard experimental work aiming to develop new treatment variants via examination of mechanisms in controlled laboratory settings. Results from experimental studies could then inform new trial arms; conversely, unexpected results found in the trial could be examined and dissected within a lab-based study. Thus, a close and rapid connection is formed between the basic and applied research, allowing more efficient bidirectional translation.

Trial parameters

Before starting the trial, the researchers need to decide on the dependent variable for comparing the arms, for example, a specific measure at a particular time point (i.e., the trial primary outcome; for example score on a measure of depression at 6 months after baseline; see the Appendix for further discussion), and the analysis method (e.g., regression, *t* test, ANCOVA etc.). Comparison of trial arms is via Bayes factors, calculation of which for these different kinds of analyses is straightforward and can be done via a range of freely available statistical packages (e.g., JASP Team, 2018; Morey & Rouder, 2015). As most commonly applied for hypothesis testing purposes, a Bayes factor quantifies the relative strength of evidence (as provided by the observed data) for the null hypothesis (e.g., no difference between means) versus an alternative hypothesis (e.g., there is a difference between means) and is essentially the probability of the observed data if one hypothesis were true divided by the probability of the observed data if the other hypothesis were true (e.g., Jeffreys, 1961; Wagenmakers, 2007). For example, if a BF of 5 for the alternative versus null hypothesis was obtained, this would indicate that the data observed were five times more probable if the alternative hypothesis were true than if the null hypothesis were true. Conversely, if the BF obtained was one fifth, this would indicate that the data observed were five times more probable if the null hypothesis were true than if the alternative hypothesis were true (for further discussion, see also e.g., Lee & Wagenmakers, 2013; Schönbrodt et al., 2017; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019).

Several further parameters (N_{min}, N_{max}, BF_{fail}, $BF_{success}$, as mentioned previously) must also be decided on to operationalize the rolling design (see Table 1). In addition to these, the prior distribution for calculation of the BFs must be specified. A full discussion of priors and their impact on BFs is beyond the scope of this article; typically, however, priors should be informed by prior knowledge about the expected gains in treatment efficiency (for further discussion of selecting priors for a sequential BF analysis design, see the Appendix of this article and Schönbrodt et al., 2017). As in a standard trial, the researchers also will need to decide how they will handle missing data (e.g., using multiple imputation or mixed models to provide an intention-to-treat approach; see the Appendix for further discussion).

The parameters for the sequential analyses will be determined by a number of factors, including pragmatic considerations (availability of participants and relevant

Table 1. Parameters for the Sequential Analyses

Parameter	Description	Influence
N_{min}	Minimum number of participants in an arm before sequential testing initiates	Smaller N_{min} allows faster dropping of an unsuccessful arm or replacement of standard arm but also increases the risk of both false positives and false negatives.
N_{max}	Maximum number of participants allowed to accrue in an arm before it is discontinued (if not reaching the superiority threshold $BF_{success}$)	N_{max} is necessary to prevent an arm continuing indefinitely if it reaches neither discontinuation threshold. A higher N_{max} gives more chance for an arm to show success against the standard arm but risks prolonging allocation of participants to a unsuccessful arm.
BF_{fail}	Threshold for a Bayes factor (BF) indicating evidence for the null hypothesis (i.e., that there is no difference between the new arm and the comparison arm).	BF_{fail} will be smaller than 1, and a smaller BF_{fail} represents a stricter threshold for rejecting the new arm. A larger (i.e., closer to 1) BF_{fail} makes it easier to reject a new arm but also increases the chance of falsely rejecting an arm that is in fact superior.
$BF_{success}$	Threshold for a Bayes factor indicating evidence for superiority of the new arm over the comparison arm.	$BF_{success}$ will be larger than 1, and a larger $BF_{success}$ represents a stricter threshold for accepting a new arm as superior to the comparison arm. A larger $BF_{success}$ thus reduces the risk of false positives but increases the risk of false negatives for any given sample size and increases the average sample size until the boundary can be reached.

resources), the relative importance given to avoiding false-positive and false-negative errors, and the kinds of effect sizes that are of interest or thought plausible. For example, if large participant numbers and time are not concerns, then stricter BF thresholds and larger minimum and maximum *N*s may be used to reduce the chances of false positives and false negatives. At an early stage of treatment development, in which only relatively substantial improvements may be of interest, or if the preference is for rapid rejection of arms not appearing to offer large superiority, a smaller N_{min} and larger (i.e., closer to 1) BF_{fail} may be appropriate. Conversely, if a treatment is well developed and small increments in average efficacy are of interest, a larger N_{min} and stricter threshold for rejection (i.e., a smaller BF_{fail} and larger N_{max} would be more useful).

If the researcher knows the kinds of effect sizes that are of interest and their relative preference for certainty versus efficiency, the likely effects of different sets of parameters can be estimated via simulated data. The Bayes Factor Design Analysis package (BFDA; Schönbrodt and Stefan, 2019; see also Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019) can be used for these purposes and allows quantification of the probabilities of false positives, false negatives, and expected overall sample sizes for a given set of parameters.

Table 2 illustrates a particular set of parameters chosen for a small to medium between-group effect size equivalent to Cohen's $d = 0.4$, as might be useful at an early stage of development in which moderate gains might still be expected or in which resource limitations mean that detection of small effects is not feasible: $N_{min} = 35$ (per arm), $N_{max} = 125$ (per arm), a BF_{fail} of one quarter, and a $BF_{success}$ of 5. The BF computation itself was performed with a directional default Cauchy prior (*rscale* parameter = $\sqrt{2}/2$).

As Table 2 illustrates, for $d = 0.4$, these parameters provide error rates similar to conventional levels used in standard trials powered to detect this effect size (hence their selection for illustration purposes). That is, the chosen parameters provide a false-positive rate (i.e., concluding $d > 0$ when $d = 0$) of 4% (equivalent to the conventional α level of < 5%) and provide 81% power to detect an efficacious training arm with a sample size of 125 when $d = 0.4$ (equivalent to the convention of 80% power, or a 20% false-negative rate). Conversely, if $d = 0$, 80% power to discontinue the arm is reached with a sample size of 75 on average, and the majority of the time (54%), the arm would be dropped at the initial analysis (i.e., at N_{min}). In most cases, the final $n = 125$ would never be reached. For example, by $n = 50$, discontinuation rates (because the BF values reached one or other threshold) range from 47% (for $d = 0.3$) to 97% (for $d = 0.8$). The associated R scripts in the Appendix of this article show how these results can be arrived at using the BFDA package.

Table 3 illustrates a different set of parameters, chosen for detecting smaller increments in effectiveness. Here, the parameters have been adjusted to be able to find a between-group effect size of $d = 0.3$, again with a standard (< 5%) false-positive rate and power (80%). Arms that are no more effective ($d = 0$) or negligibly more effective ($d = 0.1$) than the control arm are dropped on average by about 50 or 100 participants, respectively. Note that for power to find smaller effect sizes, a larger N_{min} and slightly stricter BF_{fail} are chosen; as for a standard design, a trial set up with a desired level of power to detect effect sizes of a certain magnitude will have a lower level of power to detect smaller effect sizes. Although detecting small effect sizes reliably always requires large samples regardless of the

Table 2. Illustration of Probabilities of Different Outcomes for One Example Set of Trial Parameters

"True" effect size (Cohen's *d*)	Probability of reaching threshold at each participant number (per group)									
	Discontinuation threshold					Replacement threshold				
	n = 35	*n* = 50	*n* = 75	*n* = 100	*n* = 125	*n* = 35	*n* = 50	*n* = 75	*n* = 100	*n* = 125
0 (null)	54	70	81	86	89	**1**	**3**	**3**	**4**	**4**
0.1	37	52	62	68	71	4	7	10	12	*13*
0.2	22	33	41	45	47	8	15	23	28	*32*
0.3	11	18	22	24	25	16	29	43	52	*58*
0.4	5	8	10	10	11	28	46	65	75	*81*
0.5	2	3	4	4	4	43	64	82	91	*94*
0.6	1	1	1	1	1	60	80	94	97	*98*
0.7	0	0	0	0	0	75	91	98	100	*100*
0.8	0	0	0	0	0	88	97	100	100	*100*

Note: Trial parameters: N_{min} = 35, N_{max} = 125, BF_{fail} = 1/4, and $BF_{success}$ = 5. Boldface type indicates the false-positive rate (i.e., concluding $d > 0$ when $d = 0$) as recruitment and the sequential analyses proceed, the sample size increasing from the specified N_{min} to N_{max}. The italic type indicates power to detect $d > 0$ by N_{max} at different levels of "true" effect size. *BF* = Bayes factor.

design or analytic methods chosen, the leapfrog design is relatively efficient because it allows relatively rapid discontinuation in the case of either null or medium to large effects.

The precise set of parameters a researcher selects for a particular trial and the desired false-positive and false-negative rates will be very much dependent on the specific aims and research context for the planned trial. The use of an "off-the-shelf" convenience set of parameters should therefore be avoided. In the examples above, the parameters have been chosen to provide equivalent to 80% power and an α level of < 5% for the effect size of interest because these are the conventional levels used across much of psychological research (including many clinical trials) and will be familiar to most readers. However, although such conventional thresholds may be useful (as a minimum) if the aim is to demonstrate efficacy of a particular intervention according to a standard set of criteria, if the leapfrog design is being used to develop and optimize interventions, then there is no particular a priori reason to use these rather than other thresholds. For example, if several variants of an intervention are being compared and the differences between them in practical terms (e.g., cost, time expenditure, convenience) are negligible, there may be very little practical cost associated with falsely concluding that one variant is better than another when in reality they are equivalently effective; in such circumstances, the researcher may be comfortable with a more lenient false-positive rate to reduce the chance of falsely discarding an intervention that is in fact slightly more effective. Considerations of the availability of study resources, participants, and time may also play a role: When these are more limited, the

Table 3. Illustration of Probabilities of Different Outcomes for a Different Set of Trial Parameters

"True" effect size (Cohen's *d*)	Probability of reaching threshold at each participant number (per group)									
	Discontinuation threshold					Replacement threshold				
	n = 50	*n* = 100	*n* = 150	*n* = 200	*n* = 250	*n* = 50	*n* = 100	*n* = 150	*n* = 200	*n* = 250
0 (null)	49	77	85	89	91	**1**	**3**	**3**	**4**	**4**
0.1	30	54	61	65	67	4	10	13	15	*17*
0.2	16	30	34	35	36	10	27	36	43	*48*
0.3	6	12	13	14	14	22	52	66	74	*80*
0.4	2	4	4	4	4	38	76	88	93	*95*
0.5	1	1	1	1	1	56	91	98	99	*99*

Trial parameters: N_{min} = 50, N_{max} = 250, BF_{fail} = 1/5, and $BF_{success}$ = 5. Boldface type indicates the false-positive rate (i.e., concluding $d > 0$ when $d = 0$) as recruitment and the sequential analyses proceed, the sample size increasing from the specified N_{min} to N_{max}. The italic type indicates power to detect $d > 0$ by N_{max} at different levels of "true" effect size. *BF* = Bayes factor.

researcher may wish to weigh the relative benefit of sensitivity to detect even smaller increments in effectiveness against their costs in terms of additional participants and time.

In practical terms, a researcher could start by considering what kinds of effect size increments may be worthwhile and feasible to find, the likely largest sample sizes (per arm) that would be feasible, and the minimum sample size at which they would feel comfortable discarding an arm. Pragmatic starting points for consideration of BF_{fail} and $BF_{success}$ could be conventional thresholds, for example BF > 5 as suggested by Schönbrodt et al. (2017) or BF > 10 for "strong" evidence (e.g., Lee & Wagenmakers, 2013). The researcher then can run simulations using this initial set of parameters over a range of effect sizes (e.g., using the BFDA package; Schönbrodt and Stefan, 2019; see the Appendix) to investigate the consequent predicted false-negative and false-positive rates (as in Tables 2 and 3). By iteratively adjusting the parameters and observing the effect of these adjustments over the range of plausible effect sizes, the researcher can then arrive at a constellation of trial parameters to suit their desired aims.

Cognitive training interventions as an exemplar

To provide a more concrete illustration of a potential area in which a leapfrog design may be useful, we describe a hypothetical example with reference to computerized cognitive training interventions designed to target cognitive processes implicated in psychopathology. These include approaches aimed at modifying the dysfunctional biases thought to play a role in many forms of psychopathology, termed *cognitive bias modification* (CBM) paradigms (Koster, Fox, & MacLeod, 2009), the potential clinical applications of which have been increasingly investigated (Woud & Becker, 2014). The promise of this line of research is that if such simple computerized interventions could lead to even small reductions in symptoms of psychopathology, these could be an extremely valuable addition to current treatment options. However, although there are specific cognitive training paradigms that show promise in particular target populations (e.g., approach-avoidance training to reduce relapse when administered as an addition to inpatient treatment for alcohol dependence; Wiers, Eberl, Rinck, Becker, & Lindenmeyer, 2011), on the whole, the process of clinical translation of many of the experimental laboratory paradigms has, perhaps unsurprisingly, been problematic, illustrating the challenges inherent in this process (Fox, Mackintosh, & Holmes, 2014; Koster & Bernstein, 2015).

Such cognitive training approaches provide a useful exemplar for the leapfrog design for a number of reasons. They are a clear illustration of an attempt at clinical translation within psychology research, taking an approach from theory via experimental work to clinical research, and much of this work currently sits at the translational border. Such cognitive training approaches also exemplify an area in which a close link between lab-based work and applied clinical research is particularly valuable because many lab-based tasks are unlikely to be suitable for unrestricted use in uncontrolled environments (e.g., as an Internet-delivered intervention), yet often this is what has been attempted. In addition, there are difficulties in moving between standard experimental research populations and treatment-seeking patient populations, for example, in terms of differences in motivation, likelihood and extent of change, and so on, which means that results in one particular research setting or population may not translate to another. Finally, the development of cognitive training paradigms from lab-based experimental tools to clinical interventions illustrates one pervasive problem in treatment development and optimization: Consideration of theoretical, experimental, and clinical literature may suggest a multitude of ways in which a training paradigm may be improved, and testing all the plausible options would require a prohibitive amount of time and resources—particularly if we take the conservative view that most attempted improvements are unlikely to provide a substantial incremental benefit.

For the purpose of illustration, let us assume that a group of researchers has developed a cognitive training intervention that shows consistent effects on a target cognitive process in experimental studies with healthy or subclinical participants and showed promise in initial proof-of-principle clinical studies using multiple training sessions over a short timeframe (e.g., 1 week), but for which longer lasting or more robust effects over longer time periods (e.g., a few months) have been problematic. The researchers have a number of hypotheses about refinements that might provide an advantage in terms of symptom reduction, for example, the number, duration, and scheduling of sessions; task parameters (e.g., time of presentation of stimuli, feedback given); instructions to facilitate transfer to daily life; improvements to training stimuli; and so on. The eventual aim is to develop the training into a low-intensity, self-guided, Internet-delivered intervention to reduce symptoms of depression, and the aim of the planned adaptive rolling trial is to improve outcomes over a slightly longer time period (e.g., 3 months). The resulting improved intervention would then be tested against other equivalent low-intensity interventions in a standard efficacy RCT with a longer follow-up.

Basic design. The researchers initiate a leapfrog trial with four arms: the current best version of the cognitive training and three versions that represent attempts to improve it, for example, by tweaking aspects of the schedule, instructions, or stimuli. Participants engage in the study over a 3-month period.

Measurement. In addition to baseline and outcome measures, measures of hypothesized mechanisms (e.g., the targeted cognitive process) and repeated measures of these and mood (e.g., weekly or more frequently) can be included to provide information about mechanisms and trajectories of change and inform hypotheses about improvement that could be tested in lab-based experimental work or new trial arms.

Trial parameters. For the purpose of illustration, we will use the simple outcome measure of change in symptoms of depression from baseline to the end of the trial (3 months after baseline; for discussion of issues around missing data, please refer to the Appendix). Analysis parameters will be set as $N_{min} = 35$, $N_{max} = 125$, $BF_{fail} = 1/4$, $BF_{success} = 5$, as outlined above and illustrated in Table 2. In this example, because effects of the intervention over this timeframe have not previously been evident, substantial increments in effectiveness are seen as desirable and plausible to achieve.

Maximizing real-world validity. The value of the trial is maximized by trying to provide as close as possible a real-world test of the interventions. Thus, the procedures are delivered in the way they would be in the planned end-product intervention. For example, in this case, this might mean entirely Internet-delivered, with automated reminders, no scheduled researcher contact (apart from for technical help if needed), and no financial incentives to enhance participation and adherence rates. The recruitment strategy would aim to sample those participants who would be the eventual users of the online intervention. For example, this might involve recruiting via online advertisements (e.g., search engines or social media)—the places in which people searching for help will find out about such potential interventions—but not standard research portals because these do not represent how individuals would access the intervention if eventually implemented as a self-guided Internet intervention.

Initiation of the trial. The trial is initiated with four arms (the standard arm for comparison and three experimental arms), and sequential Bayesian analyses start once arms reach N_{min}; these analyses are updated with each additional participant providing outcome data (because data collection is entirely electronic, the data extraction and analysis process could be relatively automated).

Addition of further arms. Exactly when further arms may be added is determined by a mixture of pragmatics (e.g., rate of recruitment, when other arms drop out) and availability of new ideas to test. New arms may be informed by experimental studies either from the researchers' own group or from others, results from other ongoing clinical research, and other considerations such as participant feedback. For example, an experimental study may discover that a small increase to the presentation time of the training stimuli leads to more robust training effects over a single session; it could then be tested whether using these increased stimulus presentation timings within a repeated-session administration in the trial leads to better clinical outcomes over time. Results from a parallel clinical trial from another group using a related paradigm may suggest a training schedule that looks to be particularly useful; this training schedule could then be fed into the trial. As an illustration of one further advantage of the flexibility of rolling trial designs, a previously dropped arm may be investigated within an experimental context and an improved version reentered into the trial, or previous arms may be recombined (e.g., a session of schedules with different kinds of stimuli or instructions) into new arms.

End of trial. In this example, a specific time period has been set for the leapfrog trial. As this envisaged end of the trial approaches, no new arms would be added to allow remaining arms to complete, whether by reaching the maximum participant number or dropping out beforehand. This means that the trial would end when the final remaining arm (alongside the standard arm) reaches the maximum participant number or drops out. In the case of the remaining arm reaching the maximum participant number, the Bayes factor can be used to assess whether the strength of the evidence is superior to that for the standard arm. In the case of a small to medium N_{min} (e.g., 35 in the current example), the researchers may wish to prespecify that the final arm left in the trial must reach N_{max} (or another fixed sample size) before ending the trial to avoid a situation in which an arm "wins" but has only ever been tested on a small number of participants. This also increases the statistical power for examining predictors of outcome and treatment trajectories within the final winning arm.

At the end of the trial, the winning arm is taken forward and compared against an established Internet-delivered treatment (e.g., Internet-delivered CBT) in a standard noninferiority trial with a prespecified fixed sample size. The leapfrog design therefore has been

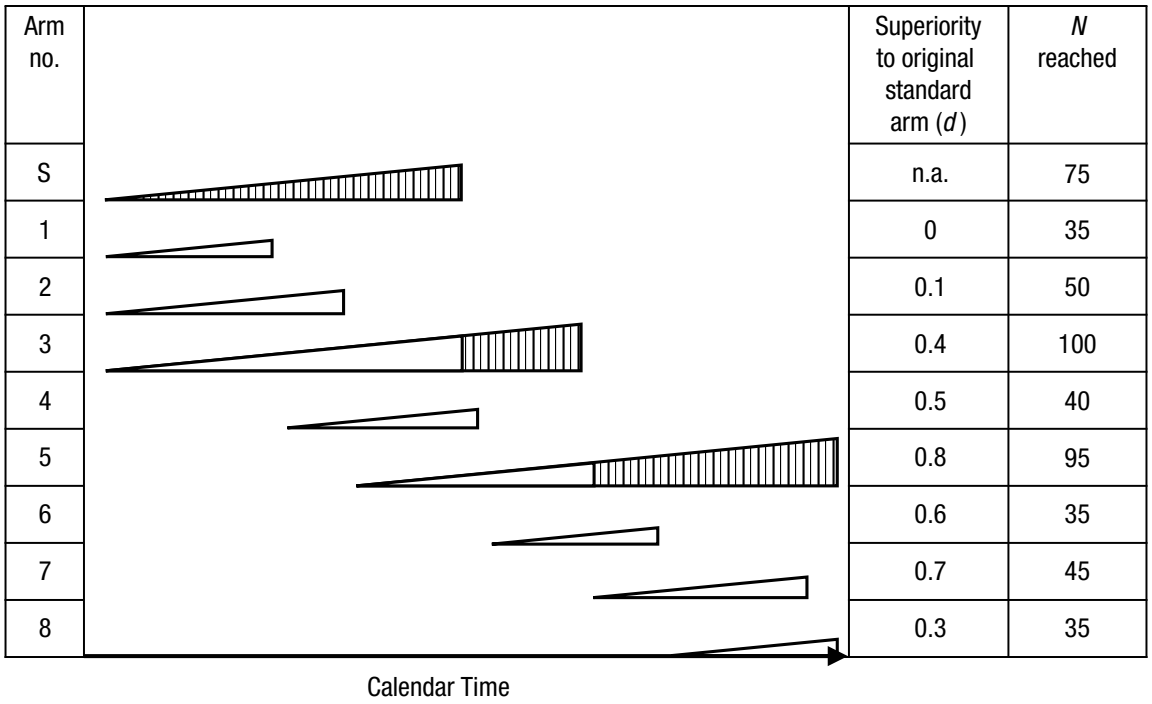

Fig. 2. Illustration of a hypothetical trial using the leapfrog design. Each triangle depicts a study arm, with the height of the triangle indicating the number of participants recruited. The standard arm is indicated by shading of the triangle, and the initial standard arm is designated S.

used to develop an optimized version of the cognitive training intervention, testing a large range of potential versions in an efficient manner, and the researchers can feel confident that the resulting version is robust enough to be worthwhile investing time and resources in future tests of clinical efficacy and effectiveness. Note that for maximal efficiency, the leapfrog trial itself could be seamlessly converted into a standard efficacy trial in a second phase (similar to the "pick a winner" design; Hills & Burnett, 2011). That is, once an arm has won the leapfrog trial, the additional treatment arm or arms for comparison could be added and enrollment started from scratch up to a prespecified fixed sample size.

Diagrammatic overview. Figure 2 illustrates how such a trial might proceed. Each triangle depicts a study arm (with the height of the triangle indicating the number of participants recruited), and the horizontal axis is calendar time. The standard arm (S) is indicated by the shaded triangle. After 75 participants, another arm (3) reaches the threshold Bayes factor to become the new standard arm. This itself is later superseded by a further arm (5). The study ends when one of the two final remaining arms drops out. Note that for arms that take over as the standard, more than N_{max} can be recruited (not illustrated in this example) because the standard arm continues recruitment until it is replaced by another. This means that randomization always includes the standard arm, and the comparison between any given intervention arm and the standard arm includes only those participants from the standard arm who could have been randomized to that comparison arm (i.e., those who were randomized while both the standard arm and the intervention arm were running—although a comparison against the full data set can also be calculated as a robustness check; for some discussion of these issues, see Hobbs et al., 2018).

Note that in this example, eight new arms have been tested against the initial standard, using a total of 510 participants. This may appear a large number of participants, even for a trial of a computerized low-intensity intervention. However, the efficiency of the leapfrog design becomes clear through comparison with standard trials. For example, even in the most efficient scenario for a classical design, using nine parallel groups in which each is compared only with the control arm, to achieve similar power as with the trial parameters illustrated in Tables 2 and 3 (i.e., 80% power at $d = 0.4$ and $\alpha < .05$), 900 participants would be needed. However, for reasons such as feasibility, a researcher using a classical design would be unlikely to embark on such a nine-arm trial. Rather, treatment-development research typically proceeds via a series of separate studies, each including a smaller number of arms. However, in the long term, this approach is extremely inefficient because of the duplication of treatment arms across trials. At the extreme, in the least efficient scenario in

which all new training variants were tested one after the other against the current best in a series of separate trials, 1,600 participants would be required when using standard trial designs. Therefore, although at first sight the example trial in Figure 2 may appear to require an unusually large number of participants, this is simply because our chosen example uses a large number of comparison arms; trials with this number of arms may be feasible only for certain kinds of interventions (e.g., Internet-delivered), and trials with fewer arms are of course possible. The efficiency of the design becomes clear when one tries to calculate how many participants (and how much time) would be needed to achieve the same outcome using standard approaches.

Of course, the relatively large increases in efficacy illustrated in this example are plausible only under certain circumstances, for example, if a potential intervention is at early stages of development with only limited efficacy, and substantial improvements would be needed to make further development worthwhile. If the aim of the trial was to improve an already established treatment, only small increments in efficacy may be plausible, and thus parameters more like those illustrated in Table 3 would be required. However, the general principle of how the trial may unfold, as illustrated in Figure 2, would remain the same.

Adaptations of the Leapfrog Design Across the Translational Process

For the purposes of introducing the basic idea of the leapfrog design, the example outlined above uses a simple superiority design at one particular stage of the translational process. However, the basic design can be adapted for a range of potential uses across the many different phases of treatment development and optimization. The following sections illustrate a selection of these, with the aim of highlighting the broader applicability of the design. For continuity with the previous section, the example of a cognitive training intervention will be used for illustration. However, the same concerns arise across all areas of psychological translational research, and some additional considerations for other kinds of interventions (e.g., face-to-face psychological therapy) will be discussed later.

Early translational investigations

If an experimental paradigm has not yet been tested in an applied setting and it is not yet clear what set of parameters may be optimal for such deployment, a trial could be set up with an initial no-treatment or monitoring-only comparison arm to establish that the new intervention is at least better than no intervention (or monitoring only). Once one treatment version has achieved this benchmark, it becomes a new comparison arm to be improved on. Notably, outperforming this active-treatment comparison arm provides a demonstration of specificity for any new arms that achieve this benchmark without having had to use potentially problematic sham or placebo arms. It might be that short time periods for measurement and the primary outcome are initially used (to allow rapid development) and then this time period can be reset and extended over time to build up longer term outcomes.

Taking a step back even further in the translational process, the leapfrog design could be applied to experimental investigations, such as single-session analogue intervention studies, or even for the development and optimization of stimuli sets or measures: If a Bayes factor can be calculated for the comparison of a relevant outcome, then the basic principles of the design can be applied. For example, in the context of developing a computerized cognitive training intervention as described above, the immediate effects of a single session of training on the cognitive outcome of interest may be investigated within an experimental design. A leapfrog design could be used to develop a version of this training that has stronger effects on this immediate outcome, which itself could then potentially be fed forward into a clinical study.

Pragmatic comparisons against a well-established treatment

Situations in which there is a well-established treatment for a particular condition but a range of alternative treatments has been developed, variations on the leapfrog design could be used to compare these fully developed interventions efficiently against the established treatment in a more pragmatic application. In such circumstances, a design more similar to the platform design of Hobbs et al. (2018) may be more suitable, retaining an established treatment throughout as the control arm. Reaching BF_{success} would then result in a new arm being discontinued for the purposes of testing in a standard preregistered fixed-N design. Hence, the design is used as a relatively resource- and time-efficient way to select potential treatments for a later more resource- and time-intensive investigation (e.g., a large pragmatic trial). Such a trial also could be set up with a superiority, noninferiority, or equivalence analysis planned (see Van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2018) or incorporate other analyses such as cost-effectiveness (see below).

Implementation as a perpetual trial

Theoretically, if the trial is being used to optimize an intervention, discover a range of potentially promising new interventions, or act as an applied testing ground

for ideas arising from basic research, it could continue indefinitely as a perpetual trial (e.g., Hobbs et al., 2018). Arms could be extracted periodically to be tested against other treatments in standard efficacy RCTs. To return to the example of a cognitive training intervention, given that such a trial would be relatively undemanding in terms of ongoing resources needed, it could run for a prolonged period of time with new training variants included as and when they are developed. However, the trial may need to be reset periodically, for example by adjusting the parameters to find smaller incremental effect size increases or switching outcome measurements if further improvement on the original primary outcome were no longer realistic. Other adjustments might be to start looking for faster trajectories of improvement or extend the timeframe for outcome measurement to build up longer lasting effects.

Beyond simple superiority designs

The example outlined in the current article assumes that a researcher is interested in simple superiority on a given parameter (e.g., change in symptoms over time) and furthermore is interested only in between-group differences of certain magnitudes. This application of the design is biased toward the status quo, requiring new arms to demonstrate a certain level of superiority to overtake the comparison arm before reaching N_{max}, and arms introduced later into the trial face a tougher test than those introduced earlier on (e.g., if an arm with a true superiority of $d = 0.4$ to the original comparison arm reaches $BF_{success}$ and becomes the new comparison arm, an arm added later on with a true superiority of $d = 0.6$ to the original arm is likely to be judged nonsuperior to the new comparison arm). If one is not interested in small (and perhaps clinically negligible) between-arm differences, this is not necessarily problematic. Further, later-introduced arms draw on a larger body of prior research (both from the performance of earlier arms in the trial and also from elsewhere) and thus may be expected to be able to achieve greater efficacy. However, in some cases, the researcher may wish to consider other options.

Adding a utility layer. Probably not all treatments have the same costs in terms of money, time, or side effects. The current example focused only on the primary outcome, conservatively staying with the previous treatment if there is no sufficient evidence that the new candidate is superior. If the new treatment is equally effective but much cheaper, however, it could be advisable to switch to the new treatment if it is noninferior. Methodologically, this could be achieved by adding a utility function to the computations that weighs costs and benefits of two competing models (Berger, 1985). Such an extended decision framework could also factor in the expected costs of (continued) experimentation and expected future gains. In the case of a cognitive training intervention, or equally for standard face-to-face psychological treatments, an example efficiency saving might be a smaller number of intervention sessions for any given level of efficacy.

Strategical order for testing multiple alternatives and dynamic adjustment of future variations. In what order are new treatments added to the leapfrog design, assuming that not all arms are started simultaneously? In many situations, the researcher may add certain arms later simply because the evidence to suggest they may be worth testing was not available at the start of the trial (if it is a trial taking place over a time period of years). However, at the start of the trial, the researcher already may have a large number of arms planned to be tested over the course of trial. Biases stemming from researcher decisions may be reduced via introducing the arms in a random order, but this might not be the most efficient way. Furthermore, knowledge gained from the early arms can be used to inform the design of later arms. Here we outline three possibilities. Although the technical details are beyond the scope of this introductory article, we hope that by mentioning these possibilities here, we can further illustrate the potential of the leapfrog design and the multiple ways in which it can be used:

1. *Explore the full treatment parameter space.* We could consider two parameters of a treatment that can be varied continuously, such as intensity of treatment (ranging from once a week to daily) and overall duration (4–12 weeks). One could test the four "corners" of the parameter space (or do a full grid search of equally spaced settings in each parameter). This gives a landscape of how treatment parameters influence the primary outcome; future tests could target those regions that are most promising.
2. *Gradually increase a treatment parameter.* In some situations, it would not be advisable to explore the full range of a parameter from the beginning. If a treatment parameter has potential side effects (e.g., the dosage in drug testing), one would probably start with a low dosage and continuously increase it while closely monitoring for potential side effects.
3. *Optimize over the treatment parameter space.* One could use well-established optimization algorithms designed to find quickly an optimum in a multidimensional parameter space (for an overview, see Varadhan, 2014). Instead of exploring the full parameter space in a grid search, such optimization quickly zooms into the most promising regions of parameter combinations by

using the full information of collected data. Issues of frequency of intervention session, or in the case of a cognitive training intervention, the number of training trials per intervention session, may be particularly well suited to this kind of approach.

Applications beyond head-to-head treatment comparisons. Whereas the current examples assume that the research is interested in head-to-head comparisons of treatments, theoretically, the unit of comparison need not be a single treatment arm. For example, in situations in which there are multiple equivalently effective treatments, it may be more clinically valuable to investigate which individuals may benefit most from which particular interventions. Increasingly sophisticated methods are being developed to attempt to make such predictions on the basis of such factors as combinations of participant baseline characteristics (e.g., Cohen & DeRubeis, 2018) or early treatment response (e.g., Forsell et al., 2019). Bayesian adaptive designs have been proposed as an efficient method for testing such matching of patients to treatments (Wason et al., 2015), and a leapfrog design could be adapted in a simple manner for efficient comparison of different treatment selection algorithms, for example, via collapsing outcomes across all treatments for each algorithm and thus having the treatment selection algorithm as the unit of comparison.

Advantages Beyond Efficiency Considerations

In addition to increased efficiency, rolling trial designs such as the leapfrog have a number of other advantages over many current methods. Because the aim is generally optimization of treatments, unless a preliminary phase is needed to establish some baseline efficacy, all participants receive active treatments. This aspect of the design brings a number of positive points. It may make it easier to recruit and retain participants and also means that engagement and efficacy are not contaminated by participants' beliefs about whether they are in a "real" or "sham" condition (Blackwell, Woud, & MacLeod, 2017). It is also ethically preferable for no participants to have to receive a sham or placebo intervention and to have relatively rapid rejection of inferior arms, as enabled by the sequential analyses. These features mean that such a design could in fact even be embedded into routine practice as every patient receives either the current best treatment or an attempt to improve on it.

For example, if an outpatient psychological therapy service wished to investigate potential low-intensity interventions for patients to engage in while on their waiting list, a rolling trial design could be used to do this testing relatively efficiently. The ability to introduce new arms while the trial is ongoing allows rapid testing of generalization of lab work to applied settings, which may save substantial research time. Mechanisms measures can be embedded into the design, and with many similar intervention variants being tested, arms potentially could be collapsed for investigation of trajectories over time. Although tests of mediation and moderation of treatment effects can be tested within a leapfrog design using the same statistical methods as in standard trial designs (e.g., Hayes & Rockwood, 2017), in a superiority design with a low N_{min}, it is likely that for many arms there will not be sufficient statistical power for robust tests of such effects. Therefore, if mechanism questions or potential subgroup analyses are a primary concern of the researcher, the sample size boundaries should be increased to ensure sufficient power. Although the primary purpose of the trial may be to optimize a treatment, a wealth of data is collected, and these data can also be reanalyzed by other researchers to address secondary research questions or make their own decisions about utility of different arms.

Cautions and Limitations

There are of course some cautions and limitations to be noted in relation to the leapfrog design and adaptive rolling trial designs more broadly. As a general point, the overall quality of a study using such an adaptive rolling design and its contribution to the scientific literature will of course depend on adoption of other good research practices to reduce bias and enhance transparency, replicability, and reproducibility (e.g., Munafò et al., 2017; Tackett et al., 2017). However, there are some additional caveats that specifically relate to the specifics of the leapfrog design.

Use of Bayesian sequential analyses

The leapfrog design makes use of Bayesian sequential analyses, with which it is possible to carry out repeated testing without compromising the overall statistical error rate. Within a NHST framework, interim decision making based on repeatedly calculated p values risks increasing the Type I error rate because it rises with each additional analysis if multiple testing is not correctly controlled for. Adaptive designs based on repeated NHST analyses therefore require ad hoc adjustments to p-value thresholds to accommodate this issue (Hobbs et al., 2018). The Bayesian analytic approach is derived from a different set of statistical principles, the nature of which means that this same problem of repeated testing does not occur; rather, with each analysis, the relative strength of evidence for the alternative versus the null hypothesis is updated according to the available data.

As the data accumulate, the evidence will likely become stronger and stronger in support of one or other of the hypotheses. This aspect of sequential Bayesian analyses has long been discussed and demonstrated more recently via simulation studies (e.g., Rouder, 2014; Schönbrodt et al., 2017). However, we note that there is some debate over the exact extent to which Bayesian analyses reduce the problems of repeated testing (for discussion of this and other caveats for sequential Bayesian analyses, see Schönbrodt et al., 2017). A potential remaining caveat to be aware of is that effect size estimates from sequential designs show a conditional bias: Early terminations either over- or underestimate the true effect, whereas late terminations underestimate it (Schönbrodt et al., 2017).

Another concern arising from repeated testing within the context of a clinical trial is whether investigator knowledge of outcomes could potentially bias results. In a standard RCT, the ideal situation is for the investigators involved in the conduct of the trial to remain blind to the trial outcomes until data collection is complete; unblinding of outcomes while the trial is ongoing can introduce a source of bias via changes (whether deliberate or unintended) in the investigators' behavior and their communications with participants. Outcome monitoring that takes place while the trial is ongoing (e.g., of potential harms) is therefore ideally conducted by a separate data monitoring committee and information shared only when absolutely necessary (e.g., for safety reasons). When using a leapfrog design, similar precautions should also be taken, with the sequential analyses conducted by individuals not involved in participant contact and other aspects of the day-to-day conduct of the trial. Automation of this aspect of the analyses and the discontinuation of arms could also help reduce this potential source of bias and may be more feasible for smaller scale research (see also Beffara, Bret, & Nalborczyk, 2018). At the very least, the nature of any interim analyses and who was aware of their results should be reported transparently in publications of the trial results, in line with the standard CONSORT recommendations for RCTs (Moher et al., 2010). However, in a sequential design, it may be that complete blinding of investigators to ongoing outcomes is not possible because the fact that a particular arm is still recruiting provides some (albeit ambiguous) information.

For these reasons, it is probably safest to view the rolling trial designs primarily as an exploratory tool for treatment development (i.e., providing a structured heuristic or decision-making process for selecting versions of a treatment that could subsequently be tested in a standard preregistered efficacy trial with a preset sample size in which outcomes are blind until data collection is complete and effect size estimates are statistically unbiased). In fact, an efficacy trial easily could be built in to the end of a rolling trial structure as a second phase (similar to the pick-a-winner design; Hills & Burnett, 2011).

Statistical analyses using small samples

Starting testing at relatively small sample sizes (if a small N_{min} is chosen) also could cause concerns because of the higher probability that arms may be unbalanced on important prognostic characteristics and the potentially higher risk of false positives or false negatives. To increase the chance that arms are balanced on important factors even with small numbers of participants, randomization strategies such as stratification and minimization can be used. In relation to the risks of false positives and false negatives, these can be quantified via simulation and reduced both via using a minimum sample size appropriate for the effect sizes of interest and adapting the parameters for stopping an arm (BF_{fail} or $BF_{success}$) appropriately. Particularly in a real-world sample, there may be much more heterogeneity among participants and in responses to interventions than in the more homogeneous samples often recruited in research, but the main effect of this additional noise in the data would be to make it harder to reach the thresholds for discontinuation (BF_{fail} or $BF_{success}$) at small sample sizes. That is, the most likely outcome of the initial analyses will be to continue data collection unless the true effect size is either in the wrong direction or huge. Thus, values of N_{min} that may appear far too small when viewed through the lens of conventional study design can be appropriate or even optimal (in terms of efficiency) within a rolling trial design.

For example, with the set of parameters presented in Table 2, $N_{min} = 35$ is chosen to be within the boundaries of conventional false-positive and false-negative rates. However, the minimum sample size in fact could be decreased further with little penalty; for example, at $N_{min} = 25$, the false-positive rate (at $d = 0$) rises only 1% to 5%, and power (at $d = 0.4$) decreases only 2% to 79%. In fact, even using a very small N_{min} of 10, in conjunction with the Bayes factor boundaries and N_{max} as in Table 2, results in only a 7% false-positive rate at $d = 0$ and provides 75% power at $d = 0.4$. Conversely, doubling N_{min} from 35 to 70 increases power (at $d = 0.4$) only 1% to 82%; although this action does halve the false-positive rate (at $d = 0$) from 4% to 2%, it comes at a high efficiency cost in terms of increased numbers of participants tested.

False positives and false negatives over time

If many different arms are tested, the likelihood of some of the results of the statistical analyses producing false

positives or false negatives is of course increased. This result, however, is not a limitation of the rolling design itself and, in fact, testing many different treatments in separate independent trials would produce a greater error rate because of the relatively economical number of comparisons (i.e., only against the comparison arm) used in the rolling trial design (see Hobbs et al., 2018). The relative efficiency of testing out a new arm within rolling trial designs such as the leapfrog, compared with setting up an entirely new trial using standard methodology, also means that the risk of discarding effective arms that by chance perform badly in the trial is lower because it is easy to reintroduce a previously dropped arm if new information suggests that it may have been dropped erroneously. Likewise, if an ineffective arm by chance manages to overtake the comparison arm, the ongoing analyses mean that it is likely that it will eventually be replaced by other arms. Thus, the design to some extent builds in safeguards against these possibilities, but as with all designs, of course it is not completely invincible to statistical flukes. Finally, in our exposition, we mostly focused on the primary outcome that is optimized. The leapfrog design is very efficient because it quickly reacts to an existing effect of superiority, which reduces the sample sizes in each arm. However, if there exist rare but severe side effects of a treatment, these relatively small sample sizes might be insufficient to detect those. Hence, the design might indicate correctly superiority of an arm concerning the primary outcome but miss a less prominent negative side effect because of the small sample size.

Evaluation of nonconcurrent trial arms

The leapfrog design provides a mechanism for continuous optimization of a treatment in an efficient manner, and the asynchronous nature of the design provides a number of advantages toward this end. These advantages include the ability to be reactive to new discoveries via the introduction of new arms as the trial progresses and the possibility of testing out a potentially large number of potential new treatment variants without needing to run all treatment arms simultaneously (which may be difficult from a practical perspective and would also result in much slower accrual of participants into each arm). However, the fact that not all treatment arms are run simultaneously limits some of the conclusions that can be drawn. For example, if treatment B outperforms treatment A with an effect size of d_{BA} and treatment C subsequently outperforms treatment B with an effect size of d_{CB}, it may be tempting to conclude that treatment C is therefore better than treatment A and even that the magnitude of this superiority is $d_{BA} + d_{CB}$. However, in the absence of a direct comparison, it is not possible to draw such conclusions with confidence.

First, it is problematic to make comparative conclusions about arms that are not run concurrently because they do not share randomization, and relative improvement may be conflated by history effects and other time-dependent factors. Second, each individual comparison is associated with a certain error rate, and these accumulate over the course of the comparisons between multiple arms (as noted in a previous section), reducing the certainty we can have about the precise ranking of arms. These problems can be reduced to some extent by increasing the sample sizes, thus reducing the error rates and increasing the potential generalizability of the findings. However, this comes at the cost of efficiency. The aim of the leapfrog design as proposed is not to achieve an overall ranking of arms but rather to improve on an initial starting point and thereafter provide a mechanism for continued improvement. If this is the aim, then the individual ranking of arms tested over the course of the trial is of lesser importance. Further, very small differences between arms likely are to be of limited practical importance. From this pragmatic perspective, it makes sense to tolerate some level of potential error in the comparisons to improve efficiency.

A third caveat to drawing comparative conclusions between multiple arms is that in a situation in which differences in the nature of the treatments being tested were substantial and obvious to participants from their descriptions, it is possible that the expectancy for improvement in any one arm may be influenced by participants' perceptions of the alternative options available. Because the arms being compared typically will all be active treatments, and often variants of a single treatment with relatively minor procedural differences, such differential expectancy effects may be reduced to some extent. That is, the potential effects on participants' expectancy of the knowledge that there is a placebo arm (see Blackwell et al., 2017; Rutherford & Roose, 2013) or even correctly guessing that they are receiving a placebo or sham treatment in a supposedly double-blind trial (e.g., Button & Munafò, 2015; Margraf et al., 1991) is avoided. However, the extent to which differential expectancy will have an influence or even occur will be unknown for any individual trial, and thus it would be useful for researchers not only to measure expectancy but also to explore whether there are any systematic changes in expectancy over the lifetime of a particular arm as the trial progresses.

Practical considerations for face-to-face psychological treatments

For illustrative purposes, a simple computerized cognitive training intervention has been used as an example through much of this article because it provides a very clear example for these purposes. For this kind of intervention and

other computerized approaches, such as Internet-delivered CBT, for which recruitment is conducted remotely and can proceed more rapidly, and for which it is relatively simple to add in a tweaked version of an intervention for comparison, it may be easier to see how a leapfrog design could be implemented. In fact, the ongoing development and evaluation of Internet-delivered variants of CBT demonstrates the vast number of parameters by which these interventions can be varied (whether in terms of content, structure, procedures, or processes targeted) to optimize treatment effects, and this demonstration in itself provides grounds for considering approaches such as the leapfrog design to investigate these possibilities more efficiently. However, much therapy is conducted face-to-face. Although an approach that increases efficiency, such as the leapfrog design, becomes even more valuable as a treatment approach becomes more time- and resource-intensive, applying the leapfrog design in such circumstances may raise some additional practical questions. The risk of bias from unblinding because of sequential analyses and how this may be addressed has been considered in the sections above, and in the following paragraphs, two further issues will be discussed.

One special consideration with regard to face-to-face therapies is the training of therapists to deliver specific therapies and subsequent monitoring of treatment fidelity. If a researcher desires to add in a slightly modified version of a treatment as a new arm partway through the trial, this of course necessitates specifying the modifications in the therapy treatment manual, training the therapists to deliver this modified therapy, and then monitoring the fidelity with which therapists do in fact deliver this therapy as intended. If the modification is relatively self-contained (e.g., addition of a specific therapeutic technique or modification of the delivery of a specific therapeutic technique), it will not necessarily be complicated, but it will cost time in preparing for a new arm and require some forward planning to avoid delays in implementing a new arm.

A second consideration is that recruitment in a face-to-face trial will generally proceed much more slowly than in a trial of a computer-delivered intervention because of limiting factors such as therapist availability and geographical restrictions. As a result, it may be feasible to test only a limited number of arms (e.g., two or three) at any one time to reach N_{min} more rapidly and start removing and replacing ineffective arms. Because of the relative expense of adding in a new face-to-face therapy arm, in terms of therapist time and other resources, the importance of earlier phase translational work, for example, mechanisms-focused experimental research and single case designs, increases in order to maximize the justification for adding in a new arm. A close reciprocal link between experimental and clinical science is seen as a key part of the further development of psychological treatments (Holmes et al., 2018) and is already illustrated in several treatment domains (e.g., exposure therapy; Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014). Given the relative speed in the possible generation of experimental findings compared with slower moving face-to-face clinical studies, there is therefore great scope for capitalizing on the asynchronous nature of the leapfrog design for these translational purposes.

However, to make optimal use of the potential efficiency savings enabled by the leapfrog design, researchers conducting trials of face-to-face therapies (and funding agencies) will have to take a longer term view of the treatment development process. Currently, treatment development typically proceeds by a series of self-contained two- or three-arm trials, many of which are likely underpowered to detect moderate but potentially clinically meaningful differences. These trials contribute to a proliferation of false-positive and false-negative results, and, given both the time taken for each individual trial to run to completion and the likely gap before the next trial starts (the intertrial lag in Fig. 1), this approach is also extremely time-inefficient. When consolidated into one leapfrog design, the associated efficiencies would allow better use of the same number of participants to provide increased power and also allow completion in an overall shorter timeframe. In settings in which an outpatient psychological therapy department is integrated within a university or other research center, this approach may be made feasible by embedding the design into this routine practice setting (as mentioned in an earlier section of the article). However, in other settings, the time required still may present challenges in view of the limited timeframes of many funding schemes and short-term nature of many academic posts. As noted in the introduction and illustrated in Figure 1, the fragmentation of clinical research into smaller chunks inadvertently reduces efficiency and increases the overall time and amount of resources needed to develop new or improved treatments. However, as the awareness and accessibility of alternative methodologies increase, clinical researchers may find it easier to convince funding agencies of the gains to be made by taking a new approach to the development and optimization of psychological therapies.

Conclusion

Although a huge amount of research effort is invested into developing or improving psychological treatments based on basic science research, this process of clinical translation is hugely challenging, and many aspects of current methods are suboptimal. This difficulty

necessitates continued work reflecting on the methods used in translational research and designing a range of improved methodologies. The development of adaptive rolling trial designs tailored for psychological treatment development, such as the leapfrog design described in the current article, holds great promise for improving the efficiency of translational research and thus the potential for improving outcomes from psychological treatments. Ultimately, the greatest savings in terms of costs and other resources can come from the application of such designs to more expensive and resource-intensive treatments such as face-to-face therapies. However, at this early stage of methodological development, suitable first testing grounds would be in contexts requiring relatively fewer resources and with faster recruitment possibilities, such as Internet-delivered interventions. This testing would enable continued development of the methodology and fuller characterization of its advantages and limitations. Such work could subsequently pave the way for wider adoption of these designs across a broader array of interventions to accelerate the development and optimization of psychological treatments.

Appendix

Further analysis decisions

As noted in the main text, in addition to deciding on the parameters for the sequential Bayesian analyses (i.e., BF_{fail}, $BF_{success}$, N_{min}, and N_{max}), a researcher planning a leapfrog trial also has to make some other analysis decisions. These decisions include handling of missing data and choosing the primary outcome for the sequential analyses and the analysis prior. Because these decisions do not present unique challenges for the leapfrog design, they are considered here rather than in the main text.

Missing data and attrition. How to handle missing data is an important consideration for all studies collecting data over time, such as clinical trials, because simply carrying out complete case analyses is likely to result in biased outcomes (e.g., Sullivan, Yelland, Lee, Ryan, & Salter, 2017). In principle, a leapfrog design is no different from a standard trial in this respect, although some special considerations are also possible. Missing data and attrition can be dealt with in a number of ways, the utility and appropriateness of which is likely to vary according to the precise circumstances of the study. One simple method is to conduct analyses on complete cases only and adjust for attrition rates in another manner. For example, participants who drop out may contribute to reaching N_{max}, such that an arm with 50% attrition would reach N_{max} while having data from only half this number of participants and is thus penalized for the high attrition rate.

Alternatively, the researcher may set an attrition threshold above which an arm is dropped regardless of apparent efficacy, or an arm may have to demonstrate no worse attrition than the control arm (via a similar Bayesian analysis) before testing efficacy. Alternatives involve analyses allowing an intention-to-treat approach, as in standard trials. For example, multiple imputation, which is frequently used in clinical trials, can be used when computing a Bayes factor when there are missing data (Hoijtink, Gu, Mulder, & Rosseel, 2019). Alternatively, mixed models can be fitted over the data (to allow for inclusion of all participants regardless of completion), and an approximate Bayes factor can be computed from statistics for the relevant fixed effect, such as a t statistic for a between-group comparison, via methods already applied or proposed in clinical research (Monden et al., 2016; Van Ravenzwaaij et al., 2018). The relative utility of these different approaches to handling missing data is likely to depend on how much missing data is expected, but simulations can be run to test the likely effects of these decision across a range of possibilities. As Bayesian analysis approaches to clinical trials become more widely used, it is likely that more detailed and standardized procedures for handling missing data in such analyses will be produced.

Primary outcome measure. In terms of the outcome variable chosen for the sequential analysis and at what timepoint this is measured, the considerations are the same as for a standard trial: This decision needs to be closely derived from the overall purpose of the trial and the research questions addressed. For example, if the aim is to improve the extent to which a treatment reduces symptoms of depression for people experiencing a depressive episode, then a relevant outcome might be a measure of depression symptoms administered at a clinically meaningful timeframe. If more immediate, or longer term, effects are of interest, then the timeframe for the outcome measurement could be adjusted accordingly. As with a standard comparative treatment trial, if two treatments with potentially different speeds of action are being compared, care must be taken to choose a timepoint for analysis that does not inadvertently induce a source of bias in favor of one or other of the treatments. For example, a short-term timepoint might favor a faster acting treatment over a slower acting but potentially more efficacious alternative. Because of the time and resource requirements of longer term follow-up, it is likely that earlier phase studies would be more likely to use shorter term outcomes, and once some indication of efficacy (e.g., effects on a target mechanisms or surrogate outcome) indicates that the greater resource investment of a longer term study would be worthwhile, this study may

be a next phase. As with standard trials, if unexpected benefits of a new treatment variant on a secondary outcome are found (or other potential advantages that are not the primary outcome, e.g., speed of improvement), although this information does not replace the primary outcome, it can be used for informing future research.

Analysis prior. The role of the analysis prior in the calculation of the Bayes factor for the sequential analyses is in the specification of the alternative hypothesis, that there is a difference between two compared treatments. The simulations and examples in this article use the "default Bayes factor" as proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009), and the prior reflects the plausibility of different Cohen's *d* effect sizes. This prior, termed a *JZS prior*, is modeled as a *t*-distribution with one degree of freedom (Cauchy distribution), the spread of which is adjusted via a scale parameter, *r*. Essentially, higher values of *r* correspond to greater plausibility of larger effect sizes, and the greater the match between the prior and the effect size in the collected data, the larger the Bayes factor in favor of the alternative hypothesis. In the absence of particularly strong evidence for the expected effect sizes, it may be simplest to use default values, such as $r = \sqrt{2}/2$, and consider safeguards such as a sensitivity analysis calculating the BF for a range of priors (Schönbrodt et al., 2017). However, potentially greater sensitivity can be acquired by tailoring the analysis prior, and the effect of changes to the prior can be examined via simulation to help planning (Stefan et al., 2019).

Simulating outcomes for a range of trial parameters using the Bayes Factor Design Analysis package in R

The probabilities for different outcomes (reaching BF_{fail} or $BF_{success}$) in Tables 2 and 3 in the main text were derived from simulations using the Bayes Factor Design Analysis (BFDA; Schönbrodt & Stefan, 2019) package for the R software environment (Version 3.6.1; R Core Team, 2019). This section of the Appendix presents a method for constructing such tables of probabilities, which would allow reproduction of the tables in the text, or simulation of potential outcomes across a range of other trial parameters. Although there are more efficient ways to produce such tables, the following process allows this without in-depth knowledge of R or the BFDA package and also allows a more in-depth examination of effects of different combinations of parameters. For a full description of the BFDA package, including installation instructions, please see https://github.com/nicebread/BFDA. The code below can also be found at https://osf.io/7rtvz/. Please check this in case there are any changes to the syntax needed to produce the required outputs.

First, simulations are run over a range of potential effect sizes. In these examples (as in the examples presented in the main text), this range is set up for an effect size corresponding to a between-group *t* test and using a directional Bayes factor. Here, results are simulated between 35 and 300 participants per condition, and 10,000 simulations are run for each effect size:

```
# install the BFDA package from Github (has to be
done only once):
install.packages("devtools"); library(devtools)
install_github("nicebread/BFDA", subdir=package")
library(BFDA)
sim.Hd00 <- BFDA.sim(expected.ES=0, type="t.
  between", n.min=35, n.max=300,
  alternative="greater", boundary=Inf, B=10000,
  verbose=TRUE)
sim.Hd01 <- BFDA.sim(expected.ES=0.1, type="t.
  between", n.min=35, n.max=300, alternative=
  "greater", boundary=Inf, B=10000, verbose=
  TRUE)
   . . .
```

up to

```
   . . .
sim.Hd08 <- BFDA.sim(expected.ES=0.8, type=
  "t.between", n.min=35, n.max=300, alternative=
  "greater", boundary=Inf, B=10000, verbose=TRUE)
```

Because these simulations can be time-consuming, it is recommended that the simulation objects be saved to allow repeated examination of the results, for example:

```
save(sim.Hd00,sim.Hd01,sim.Hd02,sim.Hd03,sim.
  Hd04,sim.Hd05,sim.Hd06,sim.Hd07,sim.
  Hd08,file="BFDAsimulationsdd.mm.yy.RData")
```

They are then reloadable via:

```
load("BFDAsimulationsdd.mm.yy.RData")
```

A useful way to examine the effects of different sets of parameters at different sample sizes (using sequential analyses) is via the BFDA plot function. For example, the following code plots (in separate windows) the percentage of simulated data sets reaching BF_{fail}, $BF_{success}$, or the N_{max} boundary without failing or succeeding across different values of N_{min} and N_{max}. These examples are for an N_{min} of 35, with a BF_{fail} of one fourth and a $BF_{success}$ of 5 (as in the upper part of Table 2). Note that the first plot goes from 35 to 36 participants to allow use of the plot function at the minimum sample size:

```
dev.new()
plot(sim.Hd00, n.min=35, n.max=36, boundary=c
  (1/4, 5))
title(main="H0, n.min=35, n.max=36,
  boundaries=1/4 and 5")
dev.new()
plot(sim.Hd00, n.min=35, n.max=50,
  boundary=c(1/4, 5))
title(main="H0, n.min=35, n.max=50,
  boundaries=1/4 and 5")
dev.new()
plot(sim.Hd00, n.min=35, n.max=75,
  boundary=c(1/4, 5))
title(main="H0, n.min=35, n.max=75,
  boundaries=1/4 and 5")
```

```
dev.new()
plot(sim.Hd00, n.min=35, n.max=100,
  boundary=c(1/4, 5))
title(main="H0, n.min=35, n.max=100,
  boundaries=1/4 and 5")
dev.new()
plot(sim.Hd00, n.min=35, n.max=125,
  boundary=c(1/4, 5))
title(main="H0, n.min=35, n.max=125,
  boundaries=1/4 and 5")
```

Figure 3 show the plots for the simulation when $d = 0$, from N_{min} ($n = 35$) to $n = 75$ (top panel) and N_{max} ($n = 125$; lower panel). By 75 participants, there already is an 81% stopping rate at BF_{fail}. By 125 participants

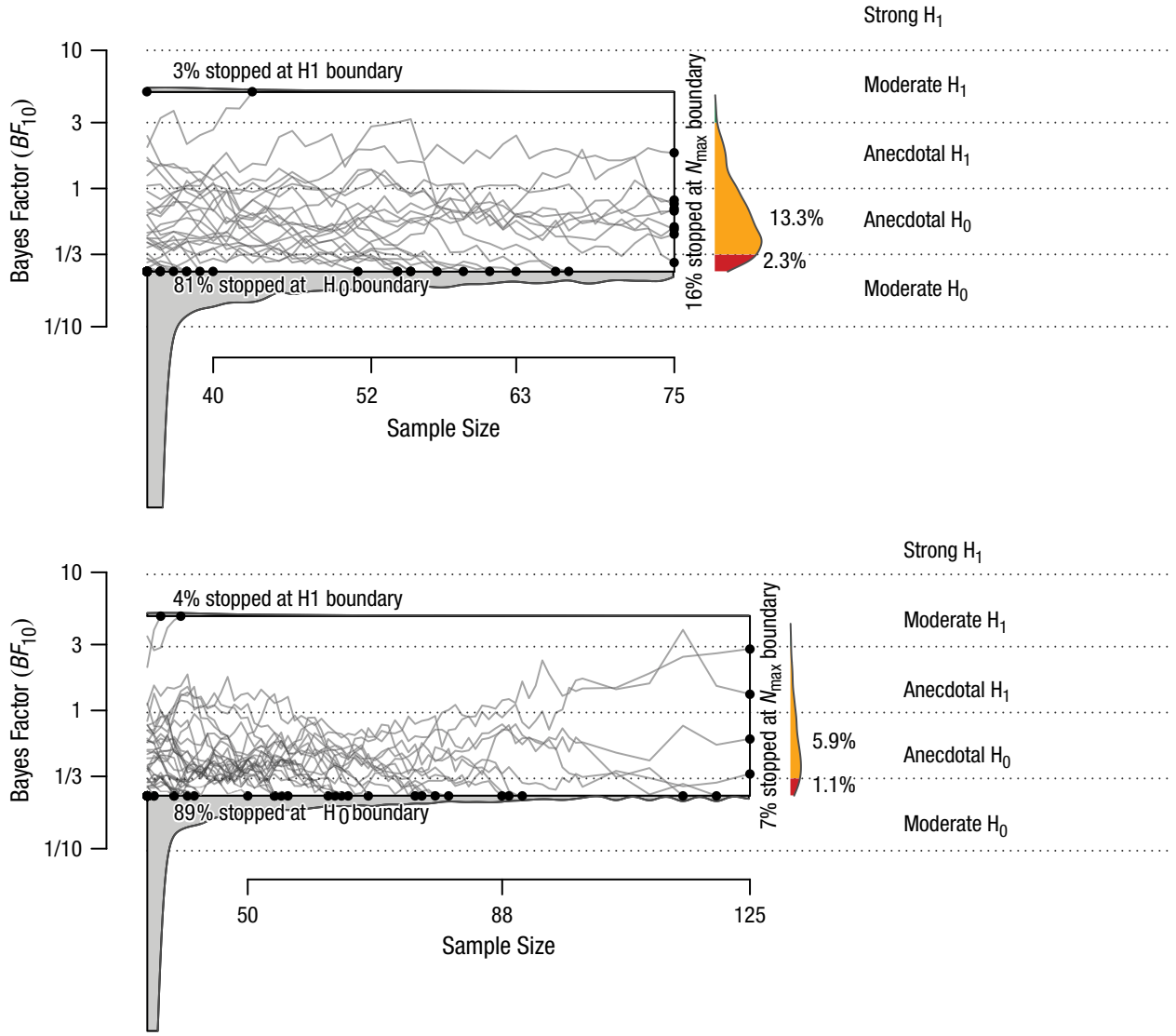

Fig. 3. Illustrative simulations from the Bayes Factor Design Analysis package, in which $d = 0$, $N_{min} = 35$, $N_{max} = 125$, $BF_{fail} = 1/4$, and $BF_{success} = 5$. The top panel shows results of sequential analyses up to $n = 75$, and the bottom panel shows results up to N_{max} ($n = 125$).

(N_{max}), this rate is now 89%, with a 4% false-positive rate (stopping at $BF_{success}$). Only 7% of the simulations remain in the "inconclusive" range by N_{max}.

Repeating the process of plotting simulated results over all the simulations for the different effect sizes allows one to see the probabilities of reaching BF_{fail} and $BF_{success}$ over a range of sample sizes for different true effect sizes. This simulation also can be repeated over a range of different parameters (for BF_{fail}, $BF_{success}$, N_{min}, and N_{max}) to investigate the impact of adjusting them. Although there are more efficient ways to obtain the probabilities as expressed in Table 2 than repeated generation of graphs, using the graphical display is helpful in providing a visual sense of how the sequential analyses unfold over the course of the (simulated) trials.

The BFDA package also allows exploration of the effect of other modifications to the trial design (e.g., different kinds of analyses, different analysis priors), and researchers can also build their own functions on top of the existing one, meaning that the planning of a trial can be very much customized to its specific research aims.

Action Editor

Stefan G. Hofmann served as action editor for this article.

Author Contributions

S. E. Blackwell conceived the leapfrog design, wrote the first draft of the manuscript, and produced the tables and figures. M. L. Woud and J. Margraf contributed to development of the ideas expressed in the article and revisions to the manuscript. F. D. Schönbrodt wrote parts of the manuscript and reviewed the code and the statistical approach. All the authors approved the final manuscript for submission.

ORCID iD

Simon E. Blackwell https://orcid.org/0000-0002-3313-7084

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

S. E. Blackwell, M. L. Woud, and J. Margraf were supported in part by the Alexander von Humboldt Professorship awarded to J. Margraf by the Alexander von Humboldt-Foundation. M. L. Woud was additionally supported by postdoctoral scholarship 32-12/4 from the Daimler and Benz Foundation and by Deutsche Forschungsgemeinschaft (DFG) Grant WO2018/2-1.

References

Andersson, G., Carlbring, P., & Lindefors, N. (2016). History and current status of ICBT. In N. Lindefors & G. Andersson (Eds.), *Guided Internet-based treatments in psychiatry* (pp. 1–16). New York, NY: Springer.

Andersson, G., & Titov, N. (2014). Advantages and limitations of Internet-based interventions for common mental disorders. *World Psychiatry*, *13*, 4–11. doi:10.1002/wps.20083

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: A (General)*, *132*, 235–244. doi:/10.2307/2343787

Beffara, B., Bret, A., & Nalborczyk, L. (2018). Automation in sequential testing: A commentary on Schönbrodt, Wagenmakers, Zehetleitner, & Perugini (2017). Retrieved from https://osf.io/mwtvk/

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York, NY: Springer.

Blackwell, S. E., Woud, M. L., & MacLeod, C. (2017). A question of control? Examining the role of control conditions in experimental psychopathology using the example of cognitive bias modification research. *Spanish Journal of Psychology, 20*, Article e54. doi:10.1017/sjp.2017.41

Button, K. S., & Munafò, M. R. (2015). Addressing risk of bias in trials of cognitive behavioral therapy. *Shanghai Archives of Psychiatry*, *27*, 144–148. doi:10.11919/j.issn.1002-0829.215042

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, *14*, 209–236. doi:10.1146/annurev-clinpsy-050817-084746

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. doi:10.1016/j.brat.2014.04.006

Elbert, T., Wilker, S., Schauer, M., & Neuner, F. (2017). Dissemination psychotherapeutischer Module für traumatisierte Geflüchtete [Dissemination of psychotherapeutic modules for traumatized refugees]. *Der Nervenarzt*, *88*, 26–33. doi:10.1007/s00115-016-0245-3

Forsell, E., Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., . . . Kaldo, V. (2019). Proof of concept for an adaptive treatment strategy to prevent failures in Internet-delivered CBT: A single-blind randomized clinical trial with insomnia patients. *American Journal of Psychiatry*, *176*, 315–323. doi:10.1176/appi.ajp.2018.18060699

Fox, E., Mackintosh, B., & Holmes, E. A. (2014). Travellers' tales in cognitive bias modification research: A commentary on the special issue. *Cognitive Therapy and Research*, *38*, 239–247. doi:10.1007/s10608-014-9604-1

Hayes, A. F., & Rockwood, N. J. (2017). Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, *98*, 39–57. doi:10.1016/j.brat.2016.11.001

Hills, R. K., & Burnett, A. K. (2011). Applicability of "pick a winner" trial design to acute myeloid leukemia. *Blood*, *118*, 2389–2394. doi:10.1182/blood-2011-02-337261

Hobbs, B. P., Chen, N., & Lee, J. J. (2018). Controlled multi-arm platform design using predictive probability. *Statistical Methods in Medical Research*, *27*, 65–78. doi:10.1177/0962280215620696

Hoijtink, H., Gu, X., Mulder, J., & Rosseel, Y. (2019). Computing Bayes factors from data with missing values. *Psychological Methods*, *24*, 253–168. doi:10.1037/met0000187

Holmes, E. A., Craske, M. G., & Graybiel, A. M. (2014). Psychological treatments: A call for mental-health science. Clinicians and neuroscientists must work together to understand and improve psychological treatments. *Nature*, *511*, 287–289. doi:10.1038/511287a

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., . . . Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, *5*, 237–286. doi:10.1016/S2215-0366(17)30513-8

JASP Team. (2018). JASP [Computer software]. Retrieved from http://jasp-stats.org/download/

Jeffreys, H. (1961). *The theory of probability*. Oxford, England: Oxford University Press.

Joyce, C. (2014). Transforming our approach to translational neuroscience: The role and impact of charitable nonprofits in research. *Neuron*, *84*, 526–532. doi:10.1016/J.NEURON.2014.10.030

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on Psychological Science*, *6*, 21–37. doi:10.1177/1745691610393527

Koster, E. H. W., & Bernstein, A. (2015). Introduction to the special issue on cognitive bias modification: Taking a step back to move forward? *Journal of Behavior Therapy and Experimental Psychiatry*, *49*, 1–4. doi:10.1016/j.jbtep.2015.05.006

Koster, E. H. W., Fox, E., & MacLeod, C. (2009). Introduction to the special section on cognitive bias modification in emotional disorders. *Journal of Abnormal Psychology*, *118*, 1–4. doi:10.1037/a0014379

Lee, W. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, England: Cambridge University Press.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A., . . . Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, *383*, 101–104. doi:10.1016/S0140-6736(13)62329-6

Margraf, J., Ehlers, A., Roth, W. T., Clark, D. B., Sheikh, J., Agras, W. S., & Taylor, C. B. (1991). How "blind" are double-blind studies? *Journal of Consulting and Clinical Psychology*, *59*, 184–187. doi:10.1037/0022-006X.59.1.184

Margraf, J., & Schneider, S. (2016). From neuroleptics to neuroscience and from Pavlov to psychotherapy: More than just the "emperor's new treatments" for mental illnesses? *EMBO Molecular Medicine*, *8*, 1115–1117. doi:10.15252/emmm.201606650

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, Article c869. doi:10.1136/bmj.c869

Monden, R., de Vos, S., Morey, R., Wagenmakers, E.-J., de Jonge, P., & Roest, A. M. (2016). Toward evidence-based medical statistics: A Bayesian analysis of double-blind placebo-controlled antidepressant trials in the treatment of anxiety disorders. *International Journal of Methods in Psychiatric Research*, *25*, 299–308. doi:10.1002/mpr.1507

Morey, R. D., & Rouder, J. N. (2015). Bayes factor: Computation of Bayes factors for common designs (Version 0.9.12-4.2) [Computer software]. Retrieved from https://cran.r-project.org/package=BayesFactor

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie, du, Sort, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, Article 0021. doi:10.1038/s41562-016-0021

Patel, V., Chowdhary, N., Rahman, A., & Verdeli, H. (2011). Improving access to psychological treatments: Lessons from developing countries. *Behaviour Research and Therapy*, *49*, 523–528. doi:10.1016/j.brat.2011.06.012

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. doi:10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:10.3758/PBR.16.2.225

Rutherford, B. R., & Roose, S. P. (2013). A model of placebo response in antidepressant clinical trials. *American Journal of Psychiatry*, *170*, 723–733. doi:10.1176/appi.ajp.2012.12040474

Schönbrodt, F. D., & Stefan, A. M. (2019). *BFDA: An R package for Bayes factor design analysis* (Version 0.5.0). Retrieved from https://github.com/nicebread/BFDA

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Stefan, A., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. (2019). A tutorial on Bayes Factor Design Analysis with informed priors. *Behavior Research Methods*, *51*, 1042–1058. doi:10.3758/s13428-018-01189-8

Sullivan, T. R., Yelland, L. N., Lee, K. J., Ryan, P., & Salter, A. B. (2017). Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clinical Trials*, *14*, 387–395. doi:10.1177/1740774517703319

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*, 742–756. doi:10.1177/1745691617690042

United Kingdom Medical Research Council. (2000). *A framework for the development and evaluation of RCTs for complex interventions to improve health*. London, England: Author.

Van Ravenzwaaij, D., Monden, R., Tendeiro, J., & Ioannidis, J. P. A. (2018). *Bayes factors for superiority, non-inferiority,*

and equivalence designs. Retrieved from http://hdl.handle.net/11370/ab5f3b34-ad4b-4a88-a7d5-7d3b8ec898a3

Varadhan, R. (2014). Numerical optimization in R: Beyond optim. *Journal of Statistical Software*, *60*, 1–3. doi:10.18637/jss.v060.i01

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wason, J. M. S., Abraham, J. E., Baird, R. D., Gournaris, I., Vallier, A.-L., Brenton, J. D., . . . Mander, A. P. (2015). A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer*, *113*, 699–705. doi:10.1038/bjc.2015.278

Wason, J. M. S., & Trippa, L. (2014). A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine*, *33*, 2206–2221. doi:10.1002/sim.6086

Watkins, E., Newbold, A., Tester-Jones, M., Javaid, M., Cadman, J., Collins, L. M., . . . Mostazir, M. (2016). Implementing multifactorial psychotherapy research in online virtual environments (IMPROVE-2): Study protocol for a phase III trial of the MOST randomized component selection method for internet cognitive-behavioural therapy for depression. *BMC Psychiatry*, *16*, Article 345. doi:10.1186/s12888-016-1054-8

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science*, *22*, 490–497. doi:10.1177/0956797611400615

Woolf, S. H. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association*, *299*, 211–213.

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*. Geneva: Author.

Woud, M. L., & Becker, E. S. (2014). Editorial for the special issue on cognitive bias modification techniques: An introduction to a time traveller's tale. *Cognitive Therapy and Research*, 38, 83–88. doi:10.1007/s10608-014-9605-0