Making the leap: From experimental psychopathology to clinical trials

Simon E Blackwell and Marcella L Woud
Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Bochum, Germany

Abstract
One important aim of experimental psychopathology research is to inform development of new interventions derived from basic science. However, testing whether a newly developed intervention is in fact effective requires moving from experimental studies to clinical trials, and this transition can pose many problems. These problems stem not only from the inherent complexity of even the simplest clinical trial but also from differences between experimental psychopathology and clinical trial research that may not always be obvious to researchers immersed in only one of these specialist areas. In this paper, we explore some of these complexities and discuss when a clinical trial may, or may not, be the best next step in the translational process. We then consider some of the ins and outs of clinical trials methodology, from design and planning through to reporting, with the aim of providing a guide for experimental psychopathology researchers thinking of making the leap from their experimental studies of mechanisms to clinical trials of novel interventions. We hope that this can help increase the chance of successful clinical translation and novel treatment development from basic science.

Keywords
Experimental psychopathology, clinical trial, clinical translation, randomized controlled trial

Date received: 28 September 2021; accepted: 25 January 2022

Many of us conducting experimental research in psychopathology hope that our work will somehow contribute to the development of new or improved psychological treatments. Although there are many ways in which experimental psychopathology research can inform our interventions (e.g. Ouimet & Ferguson, 2019; Van Den Hout et al., 2017; Waters et al., 2017), one appealingly direct route is the application of the procedures or paradigms developed and tested within our experimental work as interventions themselves within clinical settings. One such example is provided by ‘cognitive bias modification’ (CBM) procedures (Koster et al., 2009); these were initially devised to probe the causal role of cognitive biases in experimental settings (e.g. MacLeod et al., 2002; Mathews & Mackintosh, 2000), but have increasingly been investigated as clinical interventions in their own right (Blackwell, 2020; Woud & Becker, 2014). Another example comes from the line of research that started out by using visuo-spatially demanding tasks to test models of intrusive memory formation experimentally (e.g. Holmes et al., 2004), but then gradually progressed to testing the use of such tasks to reduce the occurrence of intrusive memories in real-life clinical applications (e.g. Iyadurai et al., 2018). In fact, experimental work is often heralded as showing promise for the development of novel interventions, but only a minority of this is taken forwards the next crucial clinical steps. While experimental studies are invaluable in testing theory and probing mechanisms, if we wish to draw the conclusion that a particular procedure can in fact reduce symptoms of a disorder or improve well-being, we need to test this directly in a relevant sample, over a

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

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
relevant time frame, and using relevant outcome measures—and this means conducting a clinical trial.

At face value doing a trial might seem a straightforward proposition, but in fact it is often hugely complex and problematic. The most common form of clinical trial—a randomized controlled trial (RCT) with two parallel arms—is on the surface a simple design, and this is part of its appeal. However, anyone who has been involved in the running of such a trial can attest that this superficial simplicity conceals a mass of both conceptual and practical complexity; there is in all likelihood no such thing as an ‘easy’ trial, and first-time trialists (as many experimental psychopathology researchers may be) can easily trip up over the many ‘unknown unknowns’ that litter the path from research question to trial completion.

While there are many papers and books about clinical trial methodology (e.g. Everitt & Wessely, 2008; Moher et al., 2010; Pocock, 2013), the initial transitional stage from experiment to clinical trial seems particularly problematic and therefore in need of special attention. Further, there are some specific considerations related to experimental psychopathology that articles with a broad target audience may not adequately address; this may be one reason why the design and reporting of many experimental psychopathology-derived translational studies leaves much room for improvement (see, e.g. Cristea et al., 2015). We therefore thought that it would be useful to write a paper specifically targeted at experimental psychopathology researchers without extensive trial experience who are considering making the leap from experimental study to clinical trial.

We write not as expert trialists, but rather from the perspective of experimental psychopathology researchers who have been involved in a number of clinical trials of translational interventions; there is nothing quite like arriving at the end of a trial for learning what you should have done instead. Via this paper, we aim to share some of what we have learned in this process.

We will start the paper by considering the question of whether you (the hypothetical reader) should in fact be considering a clinical trial, before exploring some of the key components of a trial using the ‘PICO’ (Patient-Intervention-Control-Outcome; e.g. Schardt et al., 2007) framework. We will then discuss two key parts of trial preparation, the protocol and the registration, before finishing with the reporting of your completed trial. We use a definition of ‘clinical trial’ broadly in line with that of the World Health Organization (e.g. 2018, p. 6), as ‘any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes’, with the additional specification that for clarity we restrict our usage of the term ‘clinical trial’ to trials in which assignment to interventions is randomized, that is, randomized controlled trials (RCTs).

**Should you really do a clinical trial?**

This section will first consider whether you—the hypothetical experimental psychopathology researcher reading this paper—should think about conducting a clinical trial, as opposed to leaving it to others (expert trialists perhaps). It will then address whether a clinical trial is in fact the best way forwards for you at this precise moment in your research.

**Should you really do a clinical trial?**

Let us imagine that you have been working on an experimental psychopathology paradigm, and sufficient, suitably robust, research has now been conducted (e.g., Craig et al., 2008; Davey, 2017) to suggest that the logical next step would be a trial to see if this paradigm could be used as an intervention to reduce symptoms of a disorder. Should you—who has never conducted a clinical trial before—really be the one to do this?

There are several reasons why you should consider conducting the trial yourself. First, if you do not do it, it will probably never happen, and the intervention will remain forever in the ‘potential promise’ phase—you are the person with the greatest motivation to take the research forwards. This is similar to what has been referred to as the ‘toothbrush problem’ in relation to psychological theories (e.g. Mischel, 2008): In the same way that most people prefer not to use someone else’s toothbrush, for a variety of reasons people often do not want to use others’ ideas for new interventions but rather develop and test their own.

Second, you likely know the paradigm inside out in a way that no-one else does, so the chance of high fidelity to the intended implementation of the paradigm is increased and the risk of a false-negative is reduced (see, e.g. Ioannidis, 2016). Of course, you will have some emotional investment in the outcome of the trial, which raises the risk of bias. You should therefore take measures to reduce potential bias, for example via robust procedures for allocation concealment, randomization and blinding (see the Supplementary Material for further discussion). However, not doing the trial at all is perhaps a little extreme: Your trial is just one step in a much longer research process that should necessarily include independent replication and larger-scale trials before any recommendations about clinical applications can be made.

Third, from a broader perspective, via planning, conducting and analysing a trial you and your colleagues will learn a huge amount about translational research and the challenges of implementing something in a clinical context. This will improve not only the quality of your own research (both clinical and experimental) but also your ability to critique, supervise, and help improve the research of others. If you have not been involved in the planning and conduct of a trial, it can be difficult to comprehend why many trials
pan out the way they do, in terms of methodological and other limitations; nothing can substitute for actually being involved in the day-to-day running and decision-making of a clinical trial in this respect.

However, while we would encourage you to consider taking the step from experimental study to clinical trial yourself, we would also very strongly suggest that you do not go it alone: Find someone who has been personally involved in running (good quality) clinical trials and ask them to give advice (e.g. feedback on your protocol) or, even better, to become a collaborator on your trial. Clinical trials are a specialism in a way that you may not fully appreciate until you have had the experience of an expert dismantling your carefully crafted trial protocol and exposing every single one of its fatal flaws. Of course, you yourself come with expertise in your research area, and there is enough dogma and unquestioned ‘tradition’ within some parts of the clinical trials world that you should not be afraid to question or reject advice given. However, by soliciting and at least considering this advice you and your trial can only benefit.

**Should you really do a clinical trial?**

In David Clark’s (2004) paper reflecting on the process of developing psychological therapies, he writes that his group would generally not conduct an RCT ‘until the treatment development work has resulted in an intervention that is associated with a substantial pre-treatment to post-treatment [within-group] effect size (greater than 1.0)’ (p. 1099).

However, a cursory glance at many trials of experimental psychopathology-derived interventions suggests that often the very first time a putative intervention meets a patient population is in the context of an RCT. If face-to-face ‘traditional’ psychological therapies require extensive development work in patient populations before moving to an RCT, why would this not be the case for experimental psychopathology-derived interventions? A premature leap from experiments to clinical trials is one of the greatest risks for false-negatives in clinical translational research. Such false-negatives can be damaging both in terms of the loss of many potentially efficacious intervention ideas, and for a field more broadly if a rash of premature null trials leads to an overall impression of failure. In this section, we will first consider reasons why there may be a tendency for premature trials in experimental psychopathology-derived clinical translation. We will then consider some particular challenges for experimental psychopathology in making the translational leap, before discussing some alternatives to RCTs as initial clinical steps (see also Dunn, O’Mahen et al., 2019). Figure 1 provides a graphical summary of some of the pressures, challenges, and risks involved, and Figure S1 in the Supplementary Material presents a version of this figure that can be annotated to help analyse your own planned trial.

**Why the rush?**. A well-conducted trial will probably consume a large amount of time, resources, and energy, and (as noted by Clark, 2004) is really only something to embark upon if the case is compelling. There are many factors that likely contribute to tendency for some experimentalists to leap prematurely to an RCT.

First, someone immersed in experimental psychology research may not be aware of the many other clinical...
research designs available besides the RCT (e.g. the huge range of single-case designs possible, or non-randomized designs such as cohort studies). Having conducted a randomized experiment within a non-clinical sample, a similar randomized study amongst patients may seem a logical next step. Researchers may also underestimate the qualitative differences between an experimental study amongst healthy (often student) participants taking part for financial or other (e.g. course credit) incentives, and a clinical trial involving people often desperately hoping for some relief from distressing and impairing symptoms (see, e.g. Field et al., 2021; Wiers et al., 2018).

Second, there are many pressures to conduct RCTs. One such pressure is a result of time and funding constraints. Cautious and methodical treatment development work may exceed the limited time frame of fixed-term research contracts, and may be difficult to get funded compared to an RCT that is easy to explain and has a greater ‘wow’ factor. Further, it may lead to publications in less ‘high impact’ journals, and may look less impressive (to some) on a CV. Another pressure, or perceived pressure, may come from the broader research community, where the necessity of large pragmatic RCTs to determine treatment effectiveness is often stressed (with good reason). However, this can lead to the impression that such RCTs, which are in fact one endpoint of a long research process, are the only valid clinical intervention studies. In turn, this can contribute to research waste via researchers viewing early-phase work as insufficiently rigorous and not worthwhile, and thus leap directly to RCTs that are inappropriate given the state of the research.

Particular challenges for experimental psychopathology-derived interventions. The developmental route to becoming a fully fledged treatment may also be much more difficult for experimental psychopathology-derived interventions compared to standard face-to-face psychological therapies, due to their lab-based rather than clinic-based gestation. In the case of a face-to-face therapeutic approach or technique, the person developing this will often be working at least part-time as a clinician, embedded within a clinical service providing psychological therapies. Initial ideas for a technique, informed by first-hand clinical experience, can often be tested within this routine service and adapted (or discarded) based on feedback or early indications of potential effectiveness. By the time the technique is tested formally for the first time, for example, in a single-case series, it will already have been well-honed, and the clinical researcher will have good reason to be confident in its likely efficacy. Often such early development work within routine practice or treatment development clinics is not published, resulting in an impression of the treatment development process that to an outsider will appear much simpler and more straightforward than is actually the case.

When it comes to developing an intervention from experimental psychopathology, the initial real-life testing lab of routine practice may not be so readily available. First, the experimental psychopathologist may not be working providing therapy, and thus may not have such ready access to a clinical setting. Second, the nature of the putative intervention may not always lend itself to integration into routine practice (e.g. if it is computer-delivered rather than simply a different clinician-delivered technique), and someone who is not clinic-based may have less of an idea what will be feasible or acceptable to patients. Experimental psychopathology-derived research will therefore often have a more difficult start, and the distance to travel before readiness for an eventual RCT will be much further.

Another challenge is moving from experimental psychopathology to clinical trials concerns measurement. While for clinical outcomes such as symptoms there are often well-established and reliable measures (e.g. self-report or clinician-administered), the immediate target of experimental psychopathology-derived interventions will often be an intermediate mechanism (e.g. a cognitive process), and the measurement of these is often much more problematic (Parsons et al., 2019). Especially for lab-derived measures relying on reaction times, there may be problems with reliability (both internal consistency and test-retest reliability), which in turn reduce sensitivity to detect change or even characterize participants at baseline. Further, when tasks have largely been developed and validated within healthy student samples (as many are, at least initially), we cannot assume that they will retain reliability and validity within clinical samples (Davey, 2017; Parsons et al., 2019), and they may not always be feasible (e.g. if a large number of experimental trials are required within the task to achieve adequate reliability, this may be a problem in clinical samples with concentration difficulties). As the assessment of changes in mechanisms often plays a central role in early-phase trials derived from experimental psychopathology, this measurement challenge can then cause particular problems.

A final, in many ways overarching, challenge in developing interventions from experimental psychopathology is that this work and the resulting putative intervention often concerns a concept or idea (e.g. targeting a specific cognitive process in a particular kind of way). However, what is tested in a trial is one specific implementation of this idea (similar to the process/procedure distinction; MacLeod & Grafton, 2016). What has been investigated or discovered in the experimental research is most likely a mechanism, or a method to probe a mechanism, that the researchers hope they can somehow leverage to have beneficial clinical effects. However, the possible parameters by which this may best (if at all) be achieved are likely to be multitudinous and complex. For example, within the area of cognitive bias modification (CBM), in which the intervention consists of individual computerized cognitive training sessions,
questions have repeatedly arisen about parameters such as the length, number, and spacing of individual training sessions; these practical parameters alone allow a huge number of possible configurations of the intervention. And while there will be a multitude of ways in which you could implement your idea (e.g. the targeting of a particular mechanism) within a treatment context, in a standard two-arm trial you can only test one specific implementation: You have just one roll of the dice. Combined with the huge array of differences between experimental studies and clinical trials, such as the nature and motivations of participants, the instructions and rationale they may receive, issues of measurement and timescale, and more (Wiers et al., 2018), the first trial can become a huge gamble: It relies on guesswork in a context where there are many unknowns, and if you guess wrong then the consequences can be far-reaching. For example, you (and others) may end up discarding the whole idea based on results of a trial in which it was the initial implementation that was sub-optimal. Premature trials, that is, those where the leap has been made without sufficient prior testing out of this guesswork, will therefore have a high rate of null results; this will perhaps particularly be the case for those modelled inappropriately on late-phase pragmatic RCTs (which have different aims compared to initial translational trials). Of course, not all of these null results will be false-negatives – many interventions are indeed ineffective, and the ideas underpinning them flawed. However, in many cases this lack of efficacy could have been discovered through suitable pre-RCT work, removing the need for a time and resource-intensive trial at all and allowing these resources to be directed elsewhere.

**Bridging designs: Single-case and pilot/feasibility studies.** If you do want to move from a healthy to a clinical population, is an RCT really the best option? As a first step it may be worth considering a single-case series design instead (e.g. Kazdin, 2019; Morley, 2017). In fact, surveying the range of psychological treatments currently evaluated as evidence-based, it is likely that all of these used single-case designs in the early stages of their development. Why should interventions derived from experimental research be the exception where such a step is not needed – particularly if they have also missed out on previous pre-research development and honing in clinical practice? Single-case series designs can provide an excellent bridge towards later-stage group-controlled designs such as RCTs. They are efficient in terms of numbers of participants, and allow testing of, for example, the schedule of interventions and measures to be used prior to a higher-risk RCT. Single-case series also provide a resource-efficient way of screening out obviously ineffective approaches; if something does not look promising in a single-case design, then it is unlikely to be worth taking forwards. Although single-case designs can be carried out to a high degree of research rigour with closely pre-specified outcomes and procedures (e.g. Dunn, Widnall et al., 2019; Hallförd et al., 2020; Khasho et al., 2019), they can also be conducted in a flexible manner for the purpose of developing and optimizing a treatment. This allows the detection of problems with a treatment (and outcome measurement) and also allows adjustments to be made from one patient in the case series to the next. Thus, by the end of the case series, you should have a well-characterized intervention that has replicated effects across a number of participants, and which can be then moved forwards to the next stage.

As an example, in our first investigation of a positive imagery-based CBM approach in a clinical (currently depressed) sample, feedback from initial participants about factors that had helped or hindered their engagement in the intervention was used to adjust the instructions and guidance we gave to subsequent participants (Blackwell & Holmes, 2010). There will be some instances where single-case approaches are not suitable, for example when a baseline phase is not possible or where other issues such as long timeframes make these designs unfeasible (e.g. it is difficult to see how the effects of approach-avoidance training on 1-year relapse rates could have been detected via a single-case series approach; Wiers et al., 2011). However, in most cases a single-case series will provide the bridge needed between experimental work and initial clinical applications.

But (you may ask), do we really need to do a single-case series to detect problems and tweak interventions prior to a formal trial – can we not achieve this simply through ‘piloting’ (however this is meant) in a patient sample? We would suggest that if you are going to ask participants to go through a research procedure, then it is best to do so in a way that generates publishable research output (e.g. via a single-case series, or even a case report or small cohort study). Publishing from the very earliest stages of treatment development makes this process much more transparent, demonstrating to others its often slow and methodical nature. It is also respectful of the time and effort that participants who may be experiencing high levels of distress and impairment have made in helping the research process. Further, if the piloting leads to the treatment approach being discarded because the chances of success seem too low, publishing this work helps reduce research waste by allowing other researchers to learn from your experience and avoid needless repetition.

Whether you conduct a formal case series or not, prior to an RCT designed to test the efficacy of your intervention you could also consider a formal feasibility or pilot trial (e.g. Lancaster et al., 2002; Leon et al., 2011); this would be a normal step before an efficacy trial within a standard psychological treatment development context, and again it is difficult to see why experimental psychopathology-derived interventions should somehow be an exception here. However, in an experimental psychopathology context such trials seem rare. One reason for this may be that the
main purpose of standard experimental psychopathology studies is testing specific hypotheses. As feasibility and pilot studies are not concerned with hypothesis-testing (apart from perhaps testing the hypothesis: ‘A full-scale trial will be feasible’) their purpose may therefore not seem clear: In a feasibility or pilot study you will not carry out statistical testing, you will generate no p-values, and you do not need to worry about statistical power to do so. In fact, you know a priori that you are underpowered – if you have sufficient power to carry out statistical tests, then it is not a pilot or feasibility study. Rather, the purpose of such studies is to test the study procedures (including e.g. recruitment, randomization, blinding, the measures you plan to use and the assessment schedule) in order to indicate whether a larger-scale trial would be feasible. Sample size recommendations therefore range from 10 per arm upwards depending on the guiding rationale (e.g. Whitehead et al., 2016). As you will collect outcome data during a feasibility or pilot trial you will generate estimates of likely efficacy of your intervention (see Hitchcock et al., 2018; Kanstrup et al., 2021; Westermann et al., 2021, for some examples); if these provide absolutely no indication of likely success, you may wish to take a step back before moving to a full-scale RCT. However, effect size estimates generated from feasibility and pilot studies should be treated with a high level of caution, particularly given the a priori knowledge that the trial is underpowered. The small sample sizes mean that any estimates are hugely imprecise, and the point estimates of effect sizes generated should generally not be used as the sole basis for sample size calculations for subsequent full-scale trials (Leon et al., 2011).

Concluding thoughts: Should you do a clinical trial? The question of when exactly your intervention is ready for a first full-scale clinical trial to test its efficacy has no straightforward answer, and at some point, the leap to such a trial needs to be made. However, it is probably safe to say that if you have no data on the effects of your intervention in the target population then an efficacy RCT is probably not the next step for you. Nevertheless, read on; the issues to be discussed below have relevance not only for such RCTs but also for earlier stages of treatment development work.

Key components of your trial

This section will consider four components of a trial that have major conceptual and practical implications, once you have decided that an RCT is the best next step. These are often summarized under the initials ‘PICO’, for Patient/Problem/Population, Intervention, Comparison, and Outcome (e.g. Huang et al., 2006; Schardt et al., 2007), and are key for specifying what questions your trial will and will not be able to address. In the following section, we will work through these four trial components in turn, focussing on aspects that we think may be particularly valuable in the context of experimental psychopathology-derived interventions. Many of these decisions to be made reflect a balancing act between the ideal real-world-relevant clinical trial and what is feasible and proportionate at an early translational stage. This is summarized in Figure 2. More detailed considerations of some of what we touch on here can be found in the Supplementary Material, as well as a version of Figure 2 (Figure S2) that can be annotated to help with the planning of your own trial.

Patients/problem/population

Who will the participants in your trial be, where will you find them, and how many do you need? In this section, we will consider several aspects of this question in relation to your sample, that is, its representativeness, how it is specified and how large it needs to be.

Representativeness of your sample. The nature of your sample will determine to what extent it may be representative of the eventual real-world end-users of your intervention, and the extent to which you can generalize from the results of your study to statements about clinical utility.

Sample characteristics: generalizability versus homogeneity. One important element of representativeness is of course the characteristics of the participants in terms of age, gender, symptom severity and so on. Here, there is often a balance to be struck between generalizability (i.e. to eventual real-world users of the intervention) and the advantages provided by having a relatively homogenous sample. The between-group effect sizes you end up calculating (and associated p-values that will often sadly determine whether people see your trial as a ‘success’ or ‘failure’) essentially derive from the ratio of the between-group variance (e.g. mean difference in pre-post change scores between two groups) to the within-group variances (or all other variance not ‘explained’ by group membership, depending on what analysis method you are using). One natural side-effect of this is that everything that increases the within-group variance (particularly within-group variance in response to the intervention) reduces your effect size and pushes your p-values further from statistical significance. Thus, given a particular sample size constraint, a more homogenous sample often provides you with more statistical power. Excluding certain complicating conditions and comorbidities, having lower and upper cut-offs for severity and other characteristics (e.g. age), and having a single recruitment source will all tend to reduce your within-group variance and, in turn, increase power. However, this comes at a cost: The more restrictions you place on your sample, the less you can generalize beyond people with these characteristics, and in fact, unrepresentativeness is a frequent complaint made about research
trials (e.g. von Wolff et al., 2014). For example, excluding people with anxiety comorbidities or suicidal ideation from a trial investigating a treatment for depression will lead to a sample that is only representative of a certain (unusual) subgroup of the population.

As with many trial design decisions, deciding on sample characteristics ends up being something of a compromise. In an early-phase trial, which is most likely what you are considering if you are reading this paper, you are probably best off aiming for a relatively homogenous sample at the expense of generalizability of your results. One example comes from the ‘proof-of-principle’ trial by Iyadurai et al. (2018), which investigated the effect of an intervention delivered in a hospital emergency department to reduce occurrence of intrusive memories following a potentially traumatic event. As this study was the first attempt to test a procedure derived from experimental psychopathology research in a clinical application, the sample was restricted to people who had suffered one particular kind of potential trauma, a road traffic accident. Keeping the sample relatively homogeneous in this way, at least in terms of trauma type, can be appropriate in this early-phase of clinical translation to increase statistical power, avoiding inappropriately large sample size requirements or otherwise high risks of false-negative results. Of course, generalizations to other kinds of trauma would need further research in which these trauma types were included, but this is not the purpose of an initial trial.

Researchers may sometimes resist placing the restrictions on participant eligibility needed to acquire a more homogeneous sample out of fear of the effects on the recruitment rate. That is, the more restrictions you place on who can take part, the greater the number of interested potential participants you will have to exclude. Given that sample sizes are one of the greatest influences on power, it may seem strange to deliberately turn away participants. However, if increasing participant numbers comes at the cost of increasing heterogeneity, these extra participants may in fact reduce rather than increase your power. As a simple (albeit extreme) illustrative example, if you were conducting a t-test on change scores, loosening your eligibility criteria in such a way that the between-group difference in change scores stayed constant but the relevant standard deviation doubled, you would need to recruit approximately four times as many participants to achieve the same level of statistical power.

What about a compromise option, using broad eligibility criteria to increase the sample size and then looking afterwards to test whether there is a particular subgroup of your participants for whom the intervention was particularly effective? Superficially, this might appear an attractive option, but in reality it is not really viable: It is highly unlikely that you will have sufficient power to identify any subgroups or moderators with any degree of reliability (e.g. Brookes et al., 2004); such power will most likely only be achieved in larger, later-phase trials (e.g. Salemink et al., 2021). An additional complication is that the more heterogeneous your sample, and the larger the sample size you then need to achieve an adequate level of statistical power, the more personnel you will need conducting research procedures and the longer your trial may take; this can then greatly increase heterogeneity of response still further and thus paradoxically reduce rather than increase your power.

Again, there is a need to compromise: At one extreme,
having one single researcher carry out all study procedures may greatly reduce ‘noise’ in the data and thus increase your statistical power (this is also true in the lab-based experimental research where researchers may vary greatly in their ability to facilitate participants’ emotional engagement in the study). However, generalizability would then be severely limited, and you would want to ascertain sooner rather than later that the intervention also worked robustly when delivered by someone other than this single individual.

Overall, at an early stage of research it is sensible to conserve resources: If you can avoid a much more expensive and longer trial with a heterogeneous sample by first finding that the intervention does not even work in a relatively homogenous sample in a smaller, less expensive trial, then your trial has been very much worthwhile. Larger trials in real-world patient samples are of course needed to demonstrate whether the intervention indeed works in such a real-world sample, but this is not where you should start.

**Recruitment source and motivation.** Participant characteristics as considered above are just one aspect of representativeness of your sample. Also important is how the participants come into contact with your intervention (including the nature of the initial contact and information provided – see the Supplementary Material for discussion), and their motivation for engaging with it. For example, you may think that in an eventual real-world application people could be directed to your intervention by their GP. However, if you then recruit your sample via convenience routes (e.g. research portals, standard advertising routes), even if you end up with a sample that is ‘representative’ of a GP-visiting sample in terms of all clinical and demographic characteristics, they are not really representative of this sample in other important ways. For example, they may differ in terms of why they are engaging in the intervention at this point in time, and their motivation to do so. As another example, unless you are intending your intervention to be something to help improve the mental health of mTurk workers who stumble across it while they search for studies to take part in, recruiting people via mTurk will not result in a sample that you can generalize to treatment-seeking populations. Conversely, advertising via Google adverts or an app store may end up with a very representative sample if this is the route via which you expect people to come across the intervention in future – but not otherwise. Most treatments are delivered to people who are seeking a treatment; a sample of people who are seeking an opportunity to take part in research (for monetary or other rewards) have a different motivation and will likely interact and engage with your intervention in a different way (see Supplementary Materials for more discussion of this, and also e.g. Field et al., 2021; Wiers et al., 2018). Of course, it may be that you do not have sufficient access to your intended eventual target audience and you adopt a compromise recruitment strategy; this is fine, but you simply need to think through the potential impact of this and be clear about the limitations of generalizability.

**Specifying your sample: Eligibility criteria.** Having decided broadly on the kind of sample you are looking for, you then need to operationalize your decisions. One crucial aspect of this, in addition to those described above, is the specific set of eligibility criteria you use.

Prior to starting your trial, you will need to specify your inclusion and exclusion criteria. Such eligibility criteria of course define the population of interest and indicate the limits of generalizability of your trial results. However, as these criteria determine who may or may not be randomized into your study, they also have quite major practical implications that are useful to think through. Regardless of how well you think you have specified your eligibility criteria, during the course of your trial, you will probably find out that there are in fact all sorts of grey areas that you have not previously considered. You can save a lot of problems by discovering as many of these grey areas as you can before you start your trial, and then clarifying them by adapting your eligibility criteria accordingly. Essentially, you want a situation in which anyone (suitably qualified) assessing the presenting participant against your eligibility criteria would come to the same conclusion about the participant’s eligibility; uncertainties lead to poorly defined and potentially biased samples, and can also cause practical problems (e.g. in explaining to participants why they are not eligible). You also want your exclusion criteria to be exhaustive, as if someone wishes to participate in your trial and technically meets your eligibility criteria, it is difficult to justify not allowing them to take part. Further, reporting a trial with a large number of ad-hoc exclusions (rightly) undermines its credibility.

One aspect of eligibility criteria that can sometimes be overlooked if the focus is on participants’ clinical characteristics is the capability of participants to complete the study procedures, even though this may seem so obvious as to be superfluous. For example, if your study procedures require completing an intervention or outcome measures over the internet, someone needs to have access to a suitable computer; if, for example, their sole opportunity to access the internet is at work via computers equipped with only Internet Explorer 7, then they will not be able to complete the study procedures. You may therefore wish to include eligibility criteria related to access to suitable technology, and assess this via (for example) requiring people to complete a procedure (e.g. an online task) with the same technical requirements as the intervention (many people may not be able to report information such as their internet browser version; of course, you can encourage people to install a different browser, if this is the issue, but it is better.
to check that this will work out prior to randomization). What if someone has pre-booked a 4-week holiday, or is moving house, 1 week after what would be the date of randomization? If this would interfere with their ability to complete study procedures, it is probably best that they are not randomized (at least at this point in time). If you do not include practical exclusion criteria that would clearly rule out these kinds of situations (and then follow-up by explicitly checking them with participants at enrolment) you may end up with many people meeting your predefined eligibility criteria for whom taking part in the study is clearly unsuitable, and this can cause problems. Therefore, it can be a good idea to overspecify, even if some eligibility criteria may appear redundant and may never be used; the extent to which any eligibility criteria were applied (or other exclusions) will in any case be reported in your final trial report.

**Sample size.** Closely tied to all other considerations of your sample is the question of how large it should be. As with many other areas of psychology, experimental psychopathology research and associated trials have tended to have a problem of insufficiently large samples sizes and low power (e.g. Rinck, 2017). Determining your sample size for an initial trial is something of a balancing act: a larger sample size generally provides greater power for statistical testing and increases the accuracy and precision of effect size estimates; conversely, given limited resources, you wish to avoid investing too much time and money in an intervention with uncertain chances of success, as this reduces the number of potential new treatments that can be tested and slows down the discovery process.

In most cases, your sample size determination will end up being a compromise between statistical power and feasibility (see Supplementary Materials for more discussion of power calculations). Of course, if you are able to easily recruit a large sample, then this obviously makes things easier. However, statistical power is not simply determined by sample size, but also by other aspects of design such as sample heterogeneity (discussed above), choice of control condition (e.g. Blackwell et al., 2017), and the outcome measure and measurement schedules used (e.g. Schuster et al., 2020). Hence, if you can increase your statistical power via adjusting the design of your trial then this can help avoid the need for unviably huge samples. However, if you can only recruit a tiny sample that provides power to detect only huge effect sizes, you probably should not be conducting such a trial at all. Rather, you may be better off using other designs (e.g. single-case series) to accumulate the evidence that might help you or others acquire funding to recruit the larger samples needed to provide worthwhile estimates of between-group effects. Further, you could focus on producing reproducible and disseminable intervention procedures (i.e. free of intellectual property, copyright, or potential commercial constraints).

Other researchers with access to greater participant numbers could then use these to take the research forwards beyond the initial development phase.

**Including experts by experience in the design process.** If you have identified the people who you hope will benefit from your new intervention, it will also be helpful if you can include representatives from this group of people in the process of planning your trial from an early stage. The involvement of such experts by experience will help increase the chance that all the different aspects of your trial, such as the intervention itself, the assessment schedule and study information, are acceptable and suitably tailored for the people you wish to engage (see, e.g. Hamilton et al., 2018 for a conceptual framework for such involvement; Dunn, O'Mahen et al., 2019 for a discussion of this in relation to psychological intervention development and Greenwood et al., in press, for an example and detailed description in relation to a specific RCT).

**Intervention(s)**

What is your intervention and how will it be implemented? This is another question that at first may seem simple to answer: surely you have been using your intervention in your experimental research and now you just need to apply it in the same way with patients? Sometimes this may be possible and appropriate, but generally, there are changes needed to adapt a procedure for clinical application and this often involves some degree of guesswork. As mentioned previously, what you may think you are testing is a particular concept or idea (e.g. using an experimental paradigm to target a particular process), but what you are really testing in your trial is (in most cases) one very specific implementation of this idea. The decisions you make about the precise implementation will have far-reaching consequences. For example, if you are unlucky and choose a poor implementation of a good idea, this may result in the idea itself (and not just the implementation) being dismissed and discarded.

**Adapting your intervention for clinical application.** There are several aspects of how exactly you might need to adapt your experimental paradigm for clinical application, and we outline these below.

**Intervention rationale, instructions and other parameters.** One aspect of the intervention that may need adaptation from an experimental setting is the rationale provided. In experimental studies, it is often the case that minimal information about a task and its aims is provided to participants, in part to reduce potential demand effects. In a clinical trial, a greater degree of explanation will probably have to be provided, including
information about potential benefits (and risks) to the participants. Without some kind of rationale, participants may not be especially motivated to engage emotionally and cognitively with the intervention in the way required for it to have an effect, and the risk of drop-out may also be higher. The exact nature of this information is likely to depend on the specifics of your trial. For example, in a very early-phase trial with a sham training comparison, detailed information about the intervention content and rationale may lead to problems with credibility of the sham condition, unbinding participants and distorting effect size estimates. Hence, in an early-phase trial the information provided might be closer to that given in experimental studies. Conversely, in a later-phase trial (e.g. where the intervention is being compared to another active treatment or treatment as usual) it would be preferable to use the instructions that would be provided to a patient in a real-world clinical application of the intervention, which would likely involve a more explicit rationale and expectancy induction.

Your experimental paradigm may also need other adaptations if the population tested in your trial differs from that in your previous experimental work. For example, if your experimental work was conducted primarily with students, but for your trial you will be recruiting patients with depression from the community, you may need to adjust the task instructions to be more comprehensible (students who take part in many studies may be familiar with some of the tasks used and unusually robust to the effects of unclear instructions). You may also need to change stimuli to be more relevant, or adjust aspects of the task timing and structure for people who may be experiencing difficulties with concentration or motivation. For example, in an RCT testing a computerized CBM approach amongst inpatients with posttraumatic stress disorder (Woud et al., 2021), initial piloting revealed that the patients found the kinds of training stimuli previously used in student samples too positive and unrelatable. We therefore adapted the training such that participants were gradually eased in to the positive training material (as per Mathews et al., 2007).

Scheduling and ‘dose’: Experimental studies often investigate the effect of a single instance or ‘dose’ of the intervention on immediate outcomes; when moving to clinical applications you may wish to investigate the effect of a larger ‘dose’ on longer-term more clinically relevant outcomes (e.g. symptom measures). The question of dosage and scheduling can then become quite complex. For example, in the context of cognitive training interventions like CBM, this means going from a single session of training to several sessions spaced out over a certain time period. The question then arises of how frequently such sessions should be completed to have noticeable effects on symptoms, and with very novel interventions, this often involves a large amount of guesswork. When we first tried a CBM intervention in the context of depression (Blackwell & Holmes, 2010), we took the approach previously used by Salemink et al. (2009) of one session every day over the course of 1 week. Part of the rationale was that by starting with this intensive training schedule we had the best chance of finding an effect, if there was one to be found. Whether daily training was in fact needed could be investigated at a later time point once there was evidence that the intervention was effective; in fact, later research also using this CBM paradigm has shown promising results using four sessions over 1 week (Pictet et al., 2016). In retrospect, engaging in the intervention daily, or almost every day (e.g. four times per week), makes sense from a number of perspectives. For example, in many domains of learning (e.g. languages and music), frequent (daily or near-daily) practice would be recommended for an optimal rate of learning, and in psychological therapies such as CBT (whether face-to-face or internet-delivered) patients engage in learning activities on a daily basis. That is, although therapy sessions themselves may typically be once per week, most days the patient will be completing therapy-related activities (e.g. records of thoughts or activities, activity-scheduling, exposure exercises or behavioural experiments); this ‘homework’ over the week between sessions is seen as key for improvement (e.g. Kazantzis et al., 2016). From this perspective, the idea that lasting change could be achieved by, for example, two 20-minute bursts behind a computer each week (which is often what has been tried – sometimes with success) seems decidedly optimistic (see also Blackwell, 2020). Following this rationale, one option for initial ‘dosing’ may then be an intensive regime to have the best chance of finding an effect (within reasonable constraints of participant burden and potential for negative effects). Later studies could then start to investigate what the optimal number of sessions (and schedule) might be (e.g. Eberl et al., 2014). Of course, the schedule must not be too burdensome or this could lead to high rates of drop-out and missing data. This again highlights the importance of earlier work such as single-case series or pilot studies to collect some data on the acceptability of the chosen intervention schedule.

Another possibility to consider as a first step might be to see whether the basic effects found in mechanisms-focussed lab-based studies (e.g. of a single ‘dose’) replicate within a clinical sample, how long they last, and to what extent they generalize (e.g. to related symptoms). You might not expect a single dose to have far-reaching effects on symptom outcomes or on related symptoms. However, even if these are your ultimate targets, there may well be a proximal mechanism on which you could detect an impact, even if only transiently. You could therefore test whether the basic effects on this proximal mechanism, as found in the lab, replicate in a clinical
sample. If so, you could then use this as the basis for investigating the optimal dosing for clinical implementation to achieve change in symptoms or other longer-term outcomes. Returning to the example of Iyadurai et al. (2018), who investigated the effect of an intervention including Tetris gameplay on intrusive memories following a road accident, in this case the intervention applied in the clinical sample was given in a similar (single) ‘dose’ to the preceding experimental studies, that is, a simple memory reactivation procedure followed by Tetris gameplay on one occasion only, within 6 hours of the accident. This therefore allowed testing of whether this specific procedure had a similar effect on the subsequent occurrence of involuntary memories when applied to real-life trauma as it had when applied to experimental analogue trauma. It would be surprising if this brief one-off ‘dose’ was the optimal implementation of this specific idea (interference with memory consolidation using a cognitively demanding task). However, replicating the effect found in experimental psychopathology research within a clinical sample provides a justification for planning further work to try to find the best way of implementing this particular idea as an intervention to maximize its clinical impact.

Theoretically, the ideal way to decide on scheduling and other dose aspects of an intervention might be to measure the effect of a single dose and how long it lasts, and titrate scheduling accordingly (as an analogy to what might be possible in the context of designing pharmacological interventions). However, in practice measuring the mechanisms targeted in a sufficiently reliable way may be too difficult, particularly given the unreliable nature of many (often behavioural) measures of mechanisms (Parsons et al., 2019) and the potential for practice effects with repeated administration. In fact, there may not be a suitably sensitive method available for measuring the proximal mechanism you think a single dose of your intervention may target in a clinical sample. This again underscores the potential value of methods such as single-case designs as an intermediate step to clinical trials, allowing you to test out your guesswork before you have to commit it to the huge time investment of an RCT.

**Adherence to the intervention.** Finally, having adapted your intervention for a clinical application, you will also want to increase the likelihood that people actually engage in it in the way required for it to have beneficial effects. However, this is a huge topic in itself and beyond the scope of a short section in this paper; a taster of some of the issues to consider here is provided in the Supplementary Material.

**Comparison/comparator**

Choosing an appropriate comparison intervention or procedure against which to compare your new intervention is often one of the hardest aspects of both experimental and early-phase clinical research (Blackwell et al., 2017). The time and mental effort investigated in this decision often feels inversely related to our enthusiasm for it: What we are really interested in is the intervention – we just know that we need something to compare it to. We will not discuss comparison or control conditions extensively here as we have done so elsewhere (Blackwell et al., 2017). A broader framework in relation to psychological therapies more generally is also provided by Gold et al. (2017). However, given that comparison conditions are so central to trial design, and present so many challenges, we think it is useful to include some discussion of them in this paper. The main point we would stress is that the comparison condition needs to be closely matched to the research question you wish to address in your study (e.g. in relation to efficacy or mechanisms). Another way of looking at this is that a specific comparison condition will generally allow you to answer one particular question; you should therefore take care to ensure that the question it answers is in fact the one you were planning to ask. The following sections consider some of the most commonly used comparison conditions.

**Waiting list control.** There is a tradition within psychological therapy research of conducting a first trial of an intervention with a waiting list as control. This has the advantage of producing a between-group effect size that *seems* easy to interpret (i.e. how effective is the intervention compared to nothing), and a waiting list is easier to implement than an active control condition. For the researcher wishing to demonstrate efficacy of their new intervention, the waitlist control also has the ‘advantage’ of almost invariably producing a statistically significant advantage in favour of the new intervention, as well as a substantial between-group effect size. However, there are also numerous problems with this traditional waitlist comparison (see also Cristea, 2019). Comparison against a waitlist will almost always produce a statistically significant result (with some rare exceptions, e.g. Lovell et al., 2017; Moritz et al., 2018), rendering the statistical testing almost meaningless. Further, the between-group effect size produced will be of doubtful utility in itself. First, there is some evidence that a waitlist can be a ‘nocebo’ condition, thus leading to overestimates of efficacy of an intervention (Furukawa et al., 2014). Second, as so many factors will influence the extent to which participants experience symptom reduction while on a waitlist (Rutherford et al., 2012), the effect size versus a waitlist achieved in any individual trial should probably not be generalized beyond that trial itself. A waitlist may be very useful as a third trial arm, for example, to verify that an established intervention is behaving as expected, or that a control condition is in fact not harmful, or for aiding inclusion in network meta-analyses, and there may be
circumstances in which it is the most appropriate comparator. However, as a single comparator, a waiting list should not be a default option, and it may do little more than establish feasibility of an intervention (in which case, a single-case series might also have done the job).

**Treatment as usual.** Theoretically, ‘treatment as usual’ (TAU) should be a tough comparator if implemented well: is your new intervention better than the current standard of care, or does it add any incremental benefit? However, the term TAU is sometimes misapplied, for example, to mean continuation of whatever treatment (or lack of it) participants are already receiving, which in fact can potentially even be a ‘nocebo’. Ideally, the TAU condition should consist of initiation of the usual treatment (e.g. someone entering an outpatient or inpatient treatment centre or visiting their GP is randomized to TAU or the new intervention). You can also compare TAU to the effect of TAU plus your intervention, that is, what is the additional benefit of adding your intervention to what people would otherwise normally receive. This provides a clinically meaningful and informative effect size. The question of specificity will of course still remain: Would adding anything to TAU also achieve similar results to your intervention? However, comparing TAU to TAU plus your intervention may provide a good starting point in situations where a TAU plus control (e.g. sham condition) comparator might be unethical (e.g. due to the study population) or unfeasible (due to likely sample size considerations).

**Sham or ‘placebo’ conditions.** A third class of comparison condition we will consider is the sham or ‘placebo’ condition. This kind of comparator has been widely used in clinical trials of experimental psychopathology-derived interventions, such as CBM, and is a natural extension of the closely matched control conditions often used in experimental work. The most commonly used sham condition is one that matches the active intervention in all aspects apart from one – the putative active ingredient (see Blackwell et al., 2017 for further elaboration). The use of such highly matched controls provides many major advantages, particularly given that within psychological therapy research more broadly such close matching is rare and often hard to achieve. However, as we have argued previously (Blackwell, 2020; Blackwell et al., 2017; see also; Becker et al., 2017; Kakoschke et al., 2017), while such a closely matched sham training condition can be hugely valuable in testing mechanisms-based questions in early-phase trials, interpretation of the between-group effect size can be problematic. This is in part because this effect size relates to the contribution of a relatively small part of a broader procedure to treatment outcomes (i.e. it is very context-dependent); it does not necessarily allow meaningful comparisons with between-group effect sizes found for other interventions, or provide a straightforward clinical interpretation. Such sham conditions are sometimes referred to as ‘placebo’ conditions, by analogy with the pill placebos used in pharmacological trials. However, this placebo analogy and terminology makes potentially problematic assumptions, (see Blackwell et al., 2017 for further discussion) and so we would suggest that it is probably best avoided. Rather, you should simply try to be very clear about what exactly you are trying to control for and how (and why).

In general, sham or ‘placebo’ comparison conditions are often appropriate and highly valuable in early-phase clinical studies over short time periods to establish specificity of the putative active ingredient of an intervention before taking it further. In designing a sham control condition, the aim should be to keep everything as close as possible to your intervention (e.g. non-specific factors such as researcher contact, expectancy and time spent engaged in an activity), and vary only that aspect (or those aspects) that comprise what you think is the ‘active ingredient’ you wish to isolate.

**Active comparators.** The final class of comparator we will consider is the active treatment comparator, that is, comparing your intervention to a specific established treatment of known efficacy or another experimental intervention. We differentiate this from TAU, as TAU generally refers to a broader package of care that may or may not include an established intervention. If there is an active treatment that is very similar in format and intensity to your intervention, this comparison can provide a similar function to a sham treatment condition, in that it can provide evidence of specificity for that aspect of your intervention that differs from the comparator (e.g. psychoeducation as a comparator for memory flexibility training; Hitchcock et al., 2018; or guided relaxation as a comparator for ‘concreteness’ training; Watkins et al., 2012). The advantage of this kind of comparison to a comparison versus sham is that it is more clinically meaningful and provides a useful effect size estimate (i.e. how much better than an existing treatment is the new intervention) in addition to demonstrating specificity; a potential disadvantage is that it may be a much tougher comparison and thus require much larger sample sizes that you may wish to use at an initial translational stage. In general, use of an active comparator allows you to start moving on to some of the most important applied clinical questions, such as whether your intervention is more effective than a potential alternative (via a superiority trial), or similarly effective but with other advantages such as cost or time commitment (via equivalence or non-inferiority trials), or even whether it is possible to identify people who will benefit more from one treatment versus another (see, e.g. Cohen & DeRubeis, 2018; Lorenzo-Luaces et al., 2021, for reviews). However, addressing these kinds of questions will need large sample sizes and will normally only be appropriate in later-phase trials once a baseline level of
efficacy for the new intervention has been established. Hence, we do not discuss active comparators further in this paper.

Coming to the end of this section on comparators, the reader in search of answers may be left somewhat dissatisfied, in that we have not provided a definite answer to this reader’s question of ‘What comparison condition should I use in my study?’ This is deliberate; we believe that the best answers to this question will come from a bottom-up consideration of the particular aims and research questions of the specific trial, and where things stand in the clinical translational process, rather than from formulaic application of simple rules. However, by working through a range of options and providing some references for further reading we hope that we have provided sufficient food for thought to help guide this thinking process. By this point in the article, the reader will not be surprised to hear that the choice of comparison condition often represents some degree of... wait for it...compromise; one specific comparison condition cannot answer every question you might find interesting, and for feasibility reasons you probably will not be able to include all the comparators required to do so. However, you can hopefully find one comparator that allows you to address one specific question; to the extent that this question is appropriate to ask at this stage of the research process, your trial will hopefully return a sensible and satisfying answer.

Outcomes

As with so many of the issues discussed so far, the issue of outcome measures may at first sight seem trivial: you know what you want to measure, don’t you? However, sadly the question of outcome measurement turns out to be not so simple after all. We will discuss outcomes first in relation to ‘Primary’ and ‘Secondary’ outcomes, and then consider some other kinds of measures you may wish or need to include. Finally, we will touch briefly on the analysis of your data.

The primary outcome measure. A key part of a clinical trial is the specification of a ‘primary outcome measure’. This measure is tied to the key question the trial aims to address, and is the one against which the ‘success’ or otherwise of your trial is judged. In most cases, your sample size calculation will be conducted to provide sufficient power to detect effects specifically in relation to your primary outcome measure.

The main purpose of pre-specifying a primary outcome measure is twofold: First, it provides some protection against the issue of multiple analyses in the context of multiple outcome measures without losing statistical power, and second, it prevents post-hoc re-rationalizations of the purpose (and success) of your trial (‘outcome switching’; Altman et al., 2017; Goldacre et al., 2019). For example, in the absence of a pre-specified primary outcome measure, someone might run a trial testing an intervention aiming to reduce symptoms of depression in which they include 20 outcome measures. Even if their intervention is no better than the comparator, there would be a good chance that one of these outcomes would provide a p-value less than .05. Theoretically, they then might then write up an article saying ‘We developed an intervention to reduce symptoms of panic, and \( p = .049 \)’.

In most cases, you should have precisely one primary outcome. Importantly, a primary outcome is defined not simply as a specific measure, but this measure at a specific time point (or if you are using a different kind of outcome, e.g. time to relapse or rate of change, then these should be defined in such a way that there is only one possible version of these). Therefore, if you are interested primarily in reducing self-report symptoms of depression, measured (for the sake of argument) using the Quick Inventory of Depressive Symptomatology (QIDS; Rush et al., 2003) you cannot simply specify ‘symptoms of depression as measured by the QIDS’ or ‘score on the QIDS’. Rather, you need to specify this more precisely, for example ‘change in symptoms of depression as measured by the QIDS from pre-to post-treatment’, or ‘scores on the QIDS at post-treatment’, according to how you eventually plan to analyse the data (see later). All other time points for this measure (e.g. the QIDS), and in fact all other outcome measures, fall under the category of secondary outcomes. Note that this means that in a situation where your treatment turns out to have delayed effects that only emerge at 6-month follow-up, your trial may technically end up being a ‘null’ trial; the 6-month follow-up effects should then be replicated in a study in which this is the primary outcome.

Your choice of primary outcome is likely to end up as a balancing act between what you think is the ultimate clinical purpose of your intervention once it has been implemented in clinical practice, and what is feasible and makes sense as a next step in the context of the current state of the research. It is easy to slip into thinking that the ‘primary’ outcome means the ‘most important’ outcome (in a clinical sense), but this may often not be appropriate in initial clinical trials. For example, perhaps you think that your intervention is targeting a mechanism that will help reduce long-term relapse rates in depression; surely then your primary outcome should be rate of relapse over a meaningful time period (e.g. 2 years)? However, an outcome such as relapse rate will often require large sample sizes to provide anywhere near sufficient statistical power, and thus your very first clinical trial will end up being huge and taking a very long time (given the follow-up period); this is not really optimal when there is no prior clinical evidence to justify such a large expenditure of time and resources.
For many initial clinical trials it may therefore be most appropriate to have as a primary outcome a measure that is meaningful and potentially predictive of the ultimate clinical outcome, but for which relatively larger between-group effect sizes may be expected. Examples might be change on a symptom outcome score (e.g. from pre to post-treatment) as opposed to a dichotomous outcome such as recovery, a measure of the key mechanism putatively targeted by your intervention (e.g. the cognitive process that you think will have downstream effects on symptoms), or an outcome that may predict the ultimate clinical outcome (e.g. in the case of relapse-prevention, reduction in residual depression symptoms). These latter kinds of outcome are sometimes referred to as intermediate or surrogate outcomes, and can potentially be problematic if misinterpreted (e.g. Heneghan et al., 2017). That is, observing an effect on the intermediate or surrogate outcome does not allow the conclusion that the intervention does in fact have an impact on the ultimate clinical outcome. However, if the results of initial trials on such intermediate outcomes look promising, this may then justify (perhaps after further optimization of the intervention’s basic effects) expending the time and resources (such as much larger participant numbers) required for a trial powered to find effects on a longer-term clinical outcome.

As an example, in the trial by Iyadurai et al. (2018), the primary outcome was number of intrusive memories of the trauma (a road traffic accident) over the subsequent week. This provided a replication in a clinical context of the basic effect found in previous experimental studies using the same outcome measure (an intrusive memory diary). Having replicated this effect, future potential options might then be: to conduct a larger trial powered to detect effects on a broader measure of posttraumatic stress disorder (PTSD) symptoms or even diagnoses at a longer time frame (i.e. to answer the question of whether the intervention could reduce future PTSD symptoms or diagnoses); to focus on the immediate effects but with a broader array of functional outcomes (i.e. does reducing intrusive memories in the first week lead to a clinically significant improvement in functioning or reduction in distress); or to focus on optimizing the potency of the intervention (e.g. by increasing the ‘dose’) prior to taking these further steps. Importantly, these kinds of studies would require much larger sample sizes, which would be difficult to justify in the absence of this initial translational step.

A caveat to the use of proximal mechanism measures as primary outcomes in initial trials (beyond the danger of over-interpreting positive results) is that if there is not already data to show that your measure is reliable, valid, and sensitive to change in a clinical sample, nominating it as your primary outcome represents a huge gamble. Of course, you can obtain some information on the suitability of a relatively untested measure in the single-case and feasibility studies that will hopefully precede your efficacy trial. However, such studies will generally involve small samples and thus can only provide imprecise estimates of a measure’s psychometric properties. On the other hand, the only way to obtain data on some of the important properties of a measure, for example, sensitivity to change and suitability as a treatment outcome, is to include the measure in a suitably large trial. From this perspective, holding off from conducting a trial until you have fully characterized the measure you want to include is not viable. You may therefore be safer using such as measure as a secondary outcome, providing you with the chance to collect the relevant psychometric data; if you do choose to use it as a primary outcome you should at least be aware of the risks. Finally, if you do choose an intermediate or surrogate primary outcome for an initial trial, the question arises of how to proceed if in a later, larger trial, effects on this surrogate do not translate to effects on the ultimate clinical outcome. To the extent that you are using a valid measure of your intermediate outcome, such a result may indicate that although the surrogate predicts the longer-term outcome, it does not play a causal role; that is, it may be a modifiable risk factor rather than a causal risk factor (Kraemer et al., 1997), and you may need to revise your theory accordingly. However, if the surrogate outcome is also a valuable clinical outcome in its own right this does not mean that the intervention should be discarded. Instead, you may need to revise how you envisage its clinical utility (e.g. accelerating recovery or providing short-term relief rather than promoting long-term changes). After all, an intervention does not need to have sustained long-term effects to be valuable, as short-term symptom relief and faster recovery can be valuable outcomes in themselves. As an analogy, paracetamol may be helpful in reducing acute fever even in cases where it has no impact on longer-term illness outcomes.

Finally, theoretically, you can choose anything as your primary outcome – it might not be a score on a questionnaire or task, but rather a model parameter (e.g. rate of change over time on a specific measure) or something more complex. For example, with the increasing interest in network approaches to disorder symptoms (Borsboom & Cramer, 2013), if you could make and operationalize a specific prediction in terms of network changes this could also provide a suitable primary outcome for a mechanisms-focused early-phase trial: you just need to be able to justify your decision and specify its operationalization.

Secondary outcomes. Secondary outcomes are easier: They are basically every outcome that is not your primary outcome. As noted above, if the measure that provides your primary outcome is used repeatedly, every administration of it is that not that defined as primary is also a secondary outcome. Secondary outcomes might include other clinical outcomes (e.g. depression in an anxiety trial and vice versa), post-randomization measures of mechanisms (see below), functional and quality of life outcomes, psychophysiological or neuronal markers and anything else you can think of. Keeping secondary outcomes to a minimum has sometimes
been advocated to reduce multiplicity. However, given the huge amount of work involved in conducting a trial and the hugely valuable opportunity they offer in investigating mechanisms of treatment, a trial with many secondary outcomes covering a wider range of domains will often ultimately be of more value (Dunn, O’Mahen et al., 2019). Further, as long as this does not compromise the primary purpose of your trial, secondary outcomes can be used as a way to build in mechanisms-focused sub-studies that have as their primary question of interest something that is very different from that of the main trial.

If you are using a short-term, intermediate or surrogate outcome as your primary outcome, you can also include longer-term outcomes or the ultimate clinical outcome of interest as secondary outcomes. In fact, if feasible, this can be hugely valuable and potentially provide estimates of the relevant effect sizes. However, you should keep in mind that as your trial was not powered to find effects on these outcomes, the statistical significance or otherwise of these secondary results is difficult to interpret, and the effect size estimates are likely to be imprecise. That is, the lack of a statistically significant difference on these ultimate clinical outcome measures does not undermine the potential promise of the intervention and should cause you no concern, because the efficacy of your intervention to have an effect on these outcome measures can only be evaluated in a suitably powered trial.

**Other measures**

**Baseline and further mechanisms measures**

Given you have gone to all the trouble of setting up and running a clinical trial, it is a good idea to make the most of this data collection opportunity (within reason). Therefore, if you can include finer-grained measures of potential mediators or moderators that might help you understand how and when your intervention works (or when it does not), then this is worth considering (if they are measured post-randomization these also count as secondary outcome measures). In fact, baseline measures collected in a clinical trial can be hugely valuable in their own right, in terms of allowing cross-sectional analyses within a clinical sample, or potentially even allowing examination of prediction or moderation of treatment outcomes (if you have adequate power; in an early-phase trial you probably do not, but this data could be invaluable in future individual patient-level meta-analyses). Finally, do not forget to also ask about participants’ views of your intervention. At the start of the trial, this involves measuring participants’ expectancy (in relation to symptom improvement), and ideally also the credibility of your intervention and the rationale provided prior to starting. At the end of the trial this involves asking participants about the acceptability of the intervention and their satisfaction with it, and any other such measures (e.g. therapeutic alliance if relevant) that may shed light on how participants perceived the intervention and the trial more broadly. In fact, the more feedback you can get from participants at the end of the trial, the better you will be able to understand the results and improve things for future trials. In an ideal world, you could consider embedding a qualitative interview study within your trial (see Darvell et al., 2015; Sánchez-Ortiz et al., 2011, for two examples), but you can at least include some open-ended feedback questions.

**Negative effects and adverse events**

A final kind of outcome to consider is how you might measure negative effects of your intervention in your trial. There has been increasing awareness that trials of psychological interventions have traditionally been poor at monitoring and reporting potential negative effects (e.g. Duggan et al., 2014; Parry et al., 2016; Rozental et al., 2014, 2018). A full discussion of negative effects and the broader category of ‘adverse events’ (i.e. any negative event that occurs to a participant in your trial regardless of whether it is or is not related to the study procedures) is beyond the scope of this paper, but the references provided above should provide a starting point for guidance. Essentially, you should pre-specify the kind of adverse (i.e. negative) events that you could potentially expect during the trial, given your sample and the study procedures, and how exactly you are going to monitor, record, and classify these (i.e. make decisions about whether they are potentially related to the interventions or not; see, e.g. Gliklich et al., 2014; Le-Rademacher et al., 2017). Monitoring and recording should cover events such as suicide attempts, clinically significant deterioration in symptoms, and patient-experienced harms (Parry et al., 2016). Information about these can potentially be derived from the outcome measures collected during the study, participant feedback (e.g. asking about any negative effects), or other data (e.g. information provided by a patient’s care team).

One welcome recent development in relation to assessment of negative effects of psychological interventions has been the construction of standardized scales to measure these (e.g. Bieda et al., 2018; Ladwig et al., 2014; Rozental et al., 2019), analogous to the measures of side-effects long used within pharmacological trials. While these scales need to be interpreted in the context of other data collected during the trial (e.g. changes in outcome measures, any records of suicidality), they offer a very useful additional tool for detecting negative effects of our interventions and it now seems difficult to justify not including such a scale in any individual trial. We of course hope to find no negative effects, but we should expect these to occur and therefore need to make the effort to record them; this information can then be used to improve the intervention and inform decisions about its relative risks and benefits.

**Analysing your outcome data.** Alongside deciding on your outcome measures, you also need to decide how you are going to analyse the data. We do not go into this in detail here (see instead the Supplementary Material), but will simply note that the concept of intention to treat will potentially make the analyses needed quite different from those you might be familiar with from your previous
experimental work. Make sure you take this into account before getting started.

Getting ready to start your trial

Having made the important design questions outlined in the previous sections of this paper, there are two final important aspects of planning for your trial that we will cover: the protocol, and the pre-registration.

The trial protocol

To a person involved in trials, it may seem odd to have to explain what a protocol is, but in fact, the term ‘protocol’ is used for many different kinds of documents across different fields of research, and even within the world of clinical trials there have been different definitions (Chan, Tetzlaff, Altman et al., 2013). Hence, it is worth disambiguating here. Basically, the protocol is a document describing the background, design, and planned conduct of your trial, from start of recruitment through data analysis to publication and dissemination of results. If you read a high-quality RCT, you will somewhere find a link to a protocol, either published as a paper, or the original uploaded in a repository somewhere, and thus there are plenty of examples around. If you are not used to writing such a document prior to starting a study, it may seem daunting and overly time-consuming. However, the process of writing the protocol is one of the most helpful in planning a trial, as it forces you to make all your ideas and plans concrete; this often reveals aspects that may be flawed or simply not feasible. The protocol is also an extremely useful way to communicate about your trial with collaborators and get feedback on the design, and all the work you put into it saves time at the later stage of analysing and publishing the results. The protocol provides the least ambiguous record of what you planned prior to starting the study, adding to the credibility of your final results and reducing potential sources of bias. You can find information as to what exactly should go into your protocol, and why, in the form of the SPIRIT (Chan, Tetzlaff, Altman et al., 2013; Chan, Tetzlaff, Gotzsche et al., 2013) guidelines. Essentially, it is a bit like the introduction and methods to what would be the eventual trial publication, plus some additional information about the management and oversight of the trial. It is this latter part (including the apparent assumption of the involvement of various committees) that perhaps has the potential to cause most doubt in the mind of the researcher as to whether their relatively modest study is really a clinical trial, but this needn’t be the case; if you are unsure what this is all about, ask someone with more trials experience, and they can hopefully demystify it for you (and you could also refer to our Supplementary Materials, where we elaborate a little on the roles and responsibilities referred to here).

How do you demonstrate that your protocol was indeed written prior to starting your trial? One simple way is to upload it somewhere like the Open Science Framework (OSF; https://osf.io) prior to the start of your study (this can be private if you do not wish it to be public right away, and on the OSF, you can also designate the protocol document as part of a registration). You could also upload the protocol onto a preprint server such as PsyArXiv (this can be done in conjunction with pre-registering it on the OSF). Additionally, you could consider publishing the protocol as a protocol paper; this comes at a cost (most journals publishing these are open access with an article processing charge), and if the main purpose of making the protocol publically available is simply public record of the plan, then not strictly necessary. However, advantages of publishing your protocol in the form of a protocol paper are that you get a publication out of it (which may be useful, particularly if there will be a long delay before the actual trial results come out), and that you will also get feedback from reviewers as to the clarity of explanation. An ideal would be to submit the protocol paper prior to starting the trial, potentially even far in advance for feedback on the design, either as a standard protocol paper or, in an ideal world, as a registered report (Chambers, 2019). However, you can also submit while a trial is recruiting (which is sometimes more feasible). One important consideration in terms of the timing is that if the protocol is publically available for the duration of the trial, participants in your trial will be able to read it (in most cases it is unlikely that they will, but it is a possibility). In some circumstances (e.g. using a sham training condition, or where demand effects may be particularly problematic), this could cause problems as it will essentially unblind participants to the intervention conditions and hypotheses. Therefore, it may not always be appropriate for your full trial protocol to be publically available while the trial is actually ongoing because it may contain information that would unblind participants or otherwise bias the trial if participants were to read it. However, there should be no reason not to upload it privately prior to trial start to obtain an independent date stamp.

Pre-registration

Regardless of what you do with your protocol, you should in all cases prospectively register your trial on a public official clinical trials registry, such as clinicaltrials.gov. This costs nothing and should be considered essential even if you also complete a registration somewhere else (e.g. OSF.io, aspredicted.org or leibniz-psychology.org). Prospective registration means submitting the registration prior to the enrolment of the very first participant; if just one person has already started your trial before you register it, technically your trial is retrospectively registered. If your country has its own clinical trials registry, you may alternatively (or additionally) wish (or be required) to register your trial there;
this can also be helpful for recruitment and national dissemination purposes. Trial registries are, of course, filled with many trial registrations, so there are plenty of examples to refer to (albeit not all of them good), but if you are unsure about what any of the items mean, ask someone.

What if you have not prospectively registered your trial, for example, completing the registration while the trial was ongoing, or after completion, or even during the review process? This can happen for a variety of reasons, mostly related to not realizing that the study you are doing will in fact be classified as a clinical trial and thus needs to be prospectively registered. The WHO guidelines (World Health Organization, 2018) are clear that lack of prospective registration should not be a barrier to publication, but that if not prospectively registered a trial should at least be retrospectively registered. And even if your registration is retrospective, you can still report your trial in a CONSORT-compliant manner: You simply have to note that the trial was not prospectively registered and provide the reason for this. One easy way to do this is when reporting your trial registration in the paper is to state when the registration took place, for example, ‘The trial was registered at clinicaltrials.gov, number NCT12345789, after the start of the trial and while recruitment was ongoing/after the end of the trial but before data analysis took place/after the end of the trial and data analysis, shortly before submission for publication of the trial results’. In terms of saying why, simply provide the reason for this in the limitations section of your discussion, for example, ‘We did not prospectively register the trial on a clinical trials registry, because at the time of planning we did not conceive of it as a clinical trial that would require such registration’ (see Horsch et al., 2017 for an example). It might feel awkward to write such a statement, but the more open you are about the limitations of your trial, the more you demonstrate your commitment to transparency and good clinical practice. Many trial publications simply write ‘The trial was registered...’ and never mention the non-prospective nature of the registration or explain it; because people will assume that this means prospective registration, such a statement is in fact misleading and a lack of acknowledgement in limitations a serious reporting issue. If someone looks at your trial registration and sees that the registration date is after the trial start date, and notes that you have not reported this (even if it is an honest mistake on your part), it will reflect badly on the trial and further raise questions about whether there are other aspects of your trial that you are not being transparent about (note: if you are reviewing a trial, always check the registration dates and make sure this is clearly addressed in the paper).

If you are in any way unsure about whether your study needs to be prospectively registered on a clinical trials registry (i.e. is it really a clinical trial?), err on the side of registering it – there is no downside. In fact, there is no good reason not to prospectively register any particular study you are planning, whether clinical or not, because the only thing prospective registration prevents you from doing is dishonestly claiming something was planned when in fact it was not. Further, journals increasingly expect or require pre-registration not only for clinical trials but also for other kinds of study. Questions about registration can only really apply to whether a clinical trials registry is appropriate or needed (as opposed to only registering somewhere like aspredicted.org or the OSF). However, given that you can also register non-randomized clinical studies such as single-case series or observational studies on clinical trial registries (and should consider doing so – particularly if an intervention is in any way involved), anything that falls broadly into the WHO clinical trials definition or involves patients or interventions in some way could potentially go here, and over-inclusivity will help avoid problems later. Note that the trial information provided on the registry is often still quite basic, so in addition you should consider publishing the protocol or some other form of more detailed registration (e.g. that includes information as to the analysis plan). And don’t forget that you will need to update your registration if anything changes, or when recruitment, and finally the trial itself, ends.

**Publishing your trial**

Having first planned then conducted your trial and analysed the outcome data, it is time to publish your results. All trial results should be published, regardless of the quality of the trial or the results found; one important purpose of the trial registration is to make a public record of the trial so the non-publication of any individual trial can be noted and followed up.

**CONSORT compliance**

Reporting of trial results is above all about transparency and awareness of the trial’s limitations, and it is here that the CONSORT guidelines (Moher et al., 2010) come in. The CONSORT guidelines are entirely about the reporting of a trial; people sometimes write of a trial that is a ‘CONSORT-compliant RCT’, but in fact, a trial itself cannot be CONSORT-compliant or otherwise – this only applies to the write-up. Theoretically, you could conduct the worst trial possible and still produce an extremely good CONSORT-compliant write-up (in fact, the CONSORT compliance of your write-up would be particularly important in this situation). Regardless of the quality of your trial, how you conducted it, the outcomes, and how you now view it with the benefit of hindsight compared to your previous idealized vision, when you come to report it, you will need to pay close attention to the CONSORT reporting guidelines. This includes not only the checklist, but the associated elaboration that explains the individual points and provides examples; often when trial reports are not fully CONSORT-
compliant it appears that people may have only looked at the checklist and therefore not quite understood what information some points are asking for. Sometimes the level of detail and transparency required may seem alien, uncomfortably revealing (of your flawed decision-making), and unnecessary. You may find your co-authors deleting (or abridging) what seem to them to be unnecessary chunks of text in an effort to reduce the word count or improve readability, but their or your opinion of the necessity of reporting these details is unimportant: If it is specified in the CONSORT guidelines, you report it. If the journal word limit does not allow reporting all of these details within the main body of your paper, put them in supplementary materials or somewhere else (e.g. with materials and data, etc., on the OSF or in another repository). Further, rather than trying to ‘spin’ a positive light on your (in hindsight, fatally flawed) trial, sweeping its imperfections under the carpet, you should revel in the transparency and (potentially) soul-baring nature of CONSORT compliance: the more transparently you can report, and the greater awareness you can demonstrate of, the limitations of your trial, the better your paper will be. If you read a lot of trial reports, particularly as a reviewer or editor, then you really notice start to notice the attention to detail and adherence to CONSORT guidelines; this immediately makes a much better impression.

And as if the basic CONSORT guidelines were not enough, there are also many extensions, such as for e-health (e.g. internet-delivered) interventions (Eysenbach & CONSORT-EHEALTH Group, 2011) or psychosocial interventions (Grant et al., 2018), and you should consult the appropriate ones. In fact, you are best off reading these before you start your trial, in the planning stage: The elaboration documents are generally very clearly written and helpful in themselves in informing good trial design and conduct, as are their reference lists. Further, if you know you will be reporting on all these minutiae of your trial, you will be more motivated to give them the thought they require. Ask yourself: what will it look like and how will I feel when I have to report this detail of my trial design in this very specific way in the final publication. If the thought makes you cringe, perhaps you need to make a change.

Data and materials availability

How can you maximize the value gained from all the time and effort that so many people – both researchers and participants – have put into making your trial happen? One of the best ways to do this is by publishing not only your trial report but also the materials used to make it happen and the individual participant-level data (see Dainer-Best et al., 2018, for a nice example). Making the materials available maximizes the chance that someone else will pick up your new intervention and have a go at taking it further forwards, helping speed up the treatment development process and offering opportunities for independent replication. Making data available allows re-use of the data for other analyses, enhances trust in the accuracy of the published results as researchers can try to reproduce the analyses (particularly if analysis scripts are also made available) and allows easy inclusion of the data in individual patient-level meta-analyses. This may be particularly important for providing sufficient sample sizes and power for examining predictors and moderators of treatment response. The data shared should ideally include item-by-item responses on questionnaires and raw data from behavioural (e.g. reaction time) tasks, as this allows the fullest re-use of the data (e.g. potentially in network analyses or to look at subgroups of symptoms, or for examining alternative ways of scoring behavioural tasks). Of course, given that clinical data from trials will be sensitive, great care must be taken to ensure that the data is as fully anonymous as possible, particularly if it is to be made open (i.e. publically available to anyone). Most of the outcome data itself (e.g. questionnaire scores, reaction times on tasks) provides no risk of identification. Whether the demographic data collected provides potential opportunities for identification of individuals will depend very much on the trial context and population; in some cases, it may be safe to make this openly available, in other cases it may be necessary to hold the demographic data (or potentially all data) in a protected repository so that it can only be provided when requested by researchers (e.g. PsyChArchives from leibniz-psychology.org/). How exactly data is shared (and materials to some extent) will depend very much on the trial itself and vary greatly; planning for sharing from the start will help this enormously. Essentially, when planning always remember that your trial is a tiny part of a much larger process; the more value for this larger process you can extract from your trial, the more worthwhile all that effort will have been.

Conclusions

There is a huge contrast between the apparent simplicity of the two-arm RCT design and the reality of actually trying to conduct such a clinical trial. While the leap from experimental psychopathology to clinical trials is necessary if the potential of experimental psychopathology for novel treatment development is to be fully realized, it presents a potential minefield of mistakes and missed opportunities. Of course, experimental psychopathology research serves many purposes, of which providing a direct route to novel treatments is only one (e.g. Becker & Rinck, 2021; Boddez et al., 2017; Ouimet et al., 2021); many experimental psychopathology researchers may have no interest in branching out to the specialist area of clinical trials, and there is no need for all to do so. However, experimental psychopathology researchers may be best placed to make these initial translational transitions from their own lab-
based work to clinical interventions. Coming to grips with trials methodology therefore opens a valuable opportunity to make this leap successfully and in doing so advance treatment development. We hope this paper leaves you (our hypothetical reader) better equipped to make the leap yourself, and perhaps even helps you to navigate the trial tightrope and arrive (ideally not too bruised and battered) in the promised land of successful clinical translation. Good luck!

Acknowledgements
We thank Amelie Requardt for her help with preparation of the figures and our daughters Mathilda and Sophie for providing us with the inspiration to create them. We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

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

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Marcella L. Woud is funded by the Deutsche Forschungsgemeinschaft (DFG) via the Emmy Noether Programme (WO 2018/3-1), a DFG research grant (WO 2018/2-1) and the SFB 1280 ‘Extinction Learning’. The funding bodies had no role in the design of the study, the collection, analysis, and interpretation of the data, or the preparation of the manuscript.

ORCID iDs
Simon E Blackwell https://orcid.org/0000-0002-3313-7084
Marcella L Woud https://orcid.org/0000-0002-4974-505X

Supplemental Material
Supplemental material for this article is available online.

Notes
1. In line with this, we will tend to draw examples from work one or other of us has been involved in – this makes it easier to comment on some of the ins and outs. Further, we will tend to draw on the same studies repeatedly as this reduces the amount of study-specific background information we have to explain in this paper.
2. Technically the WHO definition of a clinical trial can also include single-case series and cohort studies with no control group. However, our impression is that (within clinical psychology at least) people commonly use ‘clinical trial’ as another term for ‘RCT’, and thus, we feel that restricting its usage in this way (i.e. to RCTs) for this paper helps avoid potential confusion. Note that the WHO definition quoted here does not specify anything about the characteristics of the participants, for example, whether they are healthy or suffering from symptoms of a disorder, but rather emphasizes what kinds of outcome measures you are interested in. From this perspective, if you are examining the immediate effects of an experimental manipulation on, for example, state mood, a measure of cognitive processing, or a lab-based behavioural, psychophysiological or neuroimaging task, you are probably not doing a clinical trial/RCT and it might be confusing to someone reviewing the literature to designate your study as such. If you are examining the effects of this same manipulation on a symptom outcome at a later date, then it is probably safest to assume that you are now doing a clinical trial/RCT and proceed accordingly, as we will discuss towards the end of this paper.
3. If you have access to a clinical trials unit or similar organization, they will also often be very happy to advise, and any involvement or advice on their part will usually be very valuable. Further, if you are involved in a clinical trial you should try to complete Good Clinical Practice training, as this will orientate you to many of the core concepts and principles of clinical trials (and is in fact be required by some organizations).
4. Although the terms ‘pilot trial’ and ‘feasibility trial’ are often used interchangeably, formally a pilot study can be defined as a specific kind of feasibility study: ‘A feasibility study asks whether something can be done, should we proceed with it, and if so, how. A pilot study asks the same questions but also has a specific design feature: in a pilot study a future study, or part of a future study, is conducted on a smaller scale’ (Eldridge et al., 2016, p1).
5. Initial tests of a new intervention in a clinical sample (what we might also refer to as ‘early-phase’ trials), are generally ‘efficacy’ trials, that is, aiming to detect the effects of an intervention under ideal controlled conditions; in contrast, later-phase ‘effectiveness’ trials investigate ‘real-world’ performance of the intervention (e.g. Nathan et al., 2000). This efficacy/effectiveness distinction has major implications for how a trial is designed. We focus on early-phase efficacy trials in this paper.
6. Yes, this really happens.
7. Yes, this also really happens, and no, you might not find out about it (until afterwards) unless you explicitly check.

References
Altman, D. G., Moher, D., & Schulz, K. F. (2017). Harms of outcome switching in reports of randomised trials: CONSORT perspective. BMJ, 356, j396. https://doi.org/10.1136/bmj.j396
Becker, D., Jostmann, N. B., & Holland, R. W. (2017). Does approach bias modification really work in the eating domain? A commentary on Kakoschke et al. (2017). Addictive Behaviors. https://doi.org/10.1016/j.addbeh.2017.02.025
Becker, E. S., & Rinck, M. (2021). Experimental psychopathology. In Reference module in neuroscience and biobehavioral
Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLoS One*, 11(3), e0150205. https://doi.org/10.1371/journal.pone.0150205

Everitt, B. S., & Wessely, S. (2008). *Clinical trials in psychiatry*. John Wiley & Sons.

Eysenbach, G., CONSORT-EHEALTH Group. (2011). CONSORT-EHEALTH: Improving and standardizing evaluation reports of web-based and mobile health interventions. *Journal of Medical Internet Research*, 13(4), e126. https://doi.org/10.2196/jmir.192

Field, M., Christiansen, P., Hardman, C. A., Haynes, A., Jones, A., Reid, A., & Robinson, E. (2021). Translation of findings from laboratory studies of food and alcohol intake into behavior change interventions: The experimental medicine approach. *Health Psychology*, 40(12), 951–959. https://doi.org/10.1037/hea0001022

Furukawa, T. A., Noma, H., Caldwell, D. M., Honyashiki, M., Shinohara, K., Imai, H., Chen, P., Hunot, V., & Churchill, R. (2014). Waiting list may be a nocebo condition in psychotherapy trials: A contribution from network meta-analysis. *Acta Psychiatrica Scandinavica*, 130(3), 181–192. https://doi.org/10.1111/acps.12275

Gliklich, R. E., Dreyer, N. A., & Leavy, M. B. (2014). Adverse event detection, processing, and reporting. In *Registries for evaluating patient outcomes: A user's guide [internet].* (3rd ed.). Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK208615/

Gold, S. M., Enck, P., Hasselmann, H., Friele, T., Hegerl, U., Mohr, D. C., & Otte, C. (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: A decision framework. *The Lancet Psychiatry*, 4(9), 725–732. https://doi.org/10.1016/S2215-0366(17)30153-0

Goldacre, B., Drysdale, H., Marston, C., Mahtani, K. R., Dale, A., Milosevic, I., Slade, E., Hartley, P., & Heneghan, C. (2019). COMPare: Qualitative analysis of researchers’ responses to critical correspondence on a cohort of 58 misreported trials. *Trials*, 20(1), 124. https://doi.org/10.1186/s13063-019-3172-3

Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., Moher, D., Aker, J. L., Altman, D., Bhui, K., Booth, A., Clark, D., Craig, P., Eisner, M., Fraser, M. W., Gardner, F., Grant, S., Hedges, L., & Hollon, S., on behalf of the CONSORT-SPI Group. (2018). CONSORT-SPI 2018 Explanation and elaboration: Guidance for reporting social and psychological intervention trials. *Trials*, 19(1), 406. https://doi.org/10.1186/s13063-018-2735-z

Greenwood, K., Robertson, S., Vogel, E., Vella, C., Ward, T., McGourty, A., Sacadura, C., Hardy, A., Rus-Calafell, M., Collett, N., Emsley, R., Freeman, D., Fowler, D., Kuipers, E., Bebbington, P., Dunn, G., & Garety, P., the SlowMo Patient and Public Involvement (PPI) team. (in press). The impact of patient and public involvement in the SlowMo study: Reflections on peer innovation. *Health Expectations*. https://doi.org/10.1111/hex.13362

Hallford, D. J., Sharma, M. K., & Austin, D. W. (2020). Increasing anticipatory pleasure in major depression through enhancing episodic future thinking: A randomized single-case series trial. *Journal of Psychopathology and Behavioral Assessment*, 42(4), 751–764. https://doi.org/10.1007/s10862-020-09820-9

Hamilton, C. B., Hoens, A. M., Backman, C. L., McKinnon, A. M., McQuitty, S., English, K., & Li, L. C. (2018). An empirically based conceptual framework for fostering meaningful patient engagement in research. *Health Expectations*, 21(1), 396–406. https://doi.org/10.1111/hex.12635

Heneghan, C., Goldacre, B., & Mahtani, K. R. (2017). Why clinical trial outcomes fail to translate into benefits for patients. *Trials*, 18(1), 122. https://doi.org/10.1186/s13063-017-1870-2

Hitchcock, C., Gormley, S., Rees, C., Rodrigues, E., Gillard, J., Panesar, I., Wright, I. M., Hammond, E., Watson, P., Werner-Seidler, A., & Dalgleish, T. (2018). A randomised controlled trial of memory flexibility training (MemFlex) to enhance memory flexibility and reduce depressive symptomatology in individuals with major depressive disorder. *Behaviour Research and Therapy*, 110, 22–30. https://doi.org/10.1016/j.brat.2018.08.008

Holmes, E. A., Brewin, C. R., & Hennessy, R. G. (2004). Trauma films, information processing, and intrusive memory development. *Journal of Experimental Psychology: General*, 133(1), 3–22. https://doi.org/10.1037/0096-3455.133.1.3

Horsch, A., Vial, Y., Favrod, C., Harari, M. M., Blackwell, S. E., Watson, P., Iyadurai, L., Bonsall, M. B., & Holmes, E. A. (2017). Reducing intrusive traumatic memories after emergency caesarean section: A proof-of-principle randomized controlled study. *Behaviour Research and Therapy*, 94, 36–47. https://doi.org/10.1016/j.brat.2017.03.018

Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a knowledge representation for clinical questions. In AMIA Annual Symposium Proceedings, 2006, 359–363.

Ioannidis, J. P. A. (2016). Most psychotherapies do not really work, but those that might work should be assessed in biased studies. *Epidemiology and Psychiatric Sciences*, 25(05), 436–438. https://doi.org/10.1017/S2045796015000888

Iyadurai, L., Blackwell, S. E., Meiser-Stedman, R., Watson, P. C., Bonsall, M. B., Geddes, J. R., Nobre, A. C., & Holmes, E. A. (2018). Preventing intrusive memories after trauma via a brief intervention involving Tetris computer game play in the emergency department: A proof-of-concept randomized controlled trial. *Molecular Psychiatry*, 23, 674–682. https://doi.org/10.1038/mp.2017.23

Kakoschke, N., Kemps, E., & Tiggemann, M. (2017). What is the impact of social and psychological intervention trials. *Behaviour Research and Therapy*, 2018. Explanation and elaboration: Guidance for reporting social and psychological intervention trials. *Trials*, 19(1), 406. https://doi.org/10.1186/s13063-018-2735-z
computerised positive mental imagery training as a treatment adjunct in in-patient mental health settings: Randomised controlled trial. BJPsych Open, 7(6), Article e203. https://doi.org/10.1192/bjo.2021.1042

Whitehead, A. L., Julious, S. A., Cooper, C. L., & Campbell, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Statistical Methods in Medical Research, 25(3), 1057–1073. https://doi.org/10.1177/0962280215588241

Wiers, R. W., Boffo, M., & Field, M. (2018). What’s in a trial? On the importance of distinguishing between experimental lab studies and randomized controlled trials: The case of cognitive bias modification and alcohol use disorders. Journal of Studies on Alcohol and Drugs, 79(3), 333–343. https://doi.org/10.15288/jsad.2018.79.333

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(4), 490–497. https://doi.org/10.1177/0956797611400615

World Health Organization. (2018). International standards for clinical trial registries – Version 3.0. World Health Organization.

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(2), 83–88. https://doi.org/10.1007/s10608-014-9605-0

Woud, M. L., Blackwell, S. E., Shkreli, L., Würtz, F., Cwik, J. C., Margraf, J., Holmes, E. A., Steudte-Schmiedgen, S., Herrertz, S., & Kessler, H. (2021). The effects of modifying dysfunctional appraisals in posttraumatic stress disorder using a form of cognitive bias modification: Results of a randomized controlled trial in an inpatient setting. Psychotherapy and Psychosomatics. https://doi.org/10.1159/000514166