Accepted Manuscript

A commentary on research rigour in clinical psychological science: How to avoid throwing out the innovation baby with the research credibility bath water in the depression field

Barnaby D. Dunn, Heather O'Mahen, Kim Wright, Gary Brown

PII: S0005-7967(19)30103-2

DOI: https://doi.org/10.1016/j.brat.2019.103417

Article Number: 103417

Reference: BRT 103417

To appear in: *Behaviour Research and Therapy*

Received Date: 31 January 2019

Revised Date: 14 May 2019

Accepted Date: 3 June 2019

Please cite this article as: Dunn, B.D., O'Mahen, H., Wright, K., Brown, G., A commentary on research rigour in clinical psychological science: How to avoid throwing out the innovation baby with the research credibility bath water in the depression field, *Behaviour Research and Therapy* (2019), doi: https://doi.org/10.1016/j.brat.2019.103417.

A commentary on research rigour in clinical psychological science: How to avoid throwing out the innovation baby with the research credibility bath water in the depression field

Barnaby D Dunn[1]

Heather O'Mahen[1]

Kim Wright[1]

Gary Brown[2]

Mood Disorders Centre, University of Exeter, UK

Psychology Department Royal Holloway University, UK

Corresponding author: Barney Dunn, e-mail b.d.dunn@exeter.ac.uk. Mood Disorders Centre, University of Exeter, Exeter, UK, EX4 4QF.

Abstract

(161 words)

Proponents of the research credibility movement make a number of recommendations to enhance research rigour in psychology. These represent positive advances and can enhance replicability in clinical psychological science. This article evaluates whether there are any risks associated with this movement. We argue that there is the potential for research credibility principles to stifle innovation and exacerbate type II error, but only if they are applied too rigidly and beyond their intended scope by funders, journals and scientists. We outline ways to mitigate these risks. Further, we discuss how research credibility issues need to be situated within broader concerns about research waste. A failure to optimise the process by which basic science findings are used to inform the development of novel treatments (the first translational gap) and effective treatments are then implemented in real-world settings (the second translational gap) are also significant sources of research waste in depression. We make some suggestions about how to better cross these translational gaps.

In the past ten years there has been a growing recognition that many findings fail to replicate in psychology, which is in part driven by the widespread use of questionable research practices (QRPs) that inflate type I errors (i.e. false positives). A series of practical meta-science proposals have been put forward about how the rigour of psychological science can be enhanced (which can be referred to as the research credibility movement). These proposals are now being extended from the basic science arena to other domains of applied psychology including clinical psychology. This is a timely, important and useful change that may greatly benefit the field. However, as with most change, there could be risks as well as opportunities depending on how these recommendations are implemented.

This article will identify possible risks in implementing research credibility recommendations in clinical psychological science, consider how likely these are to occur, and discuss ways to mitigate against them. We use intervention development in depression as a focus to illustrate these points. We argue there is a risk that the research credibility movement may inhibit innovation and exacerbate type II error in the field if its principles are applied in a rigid, "one size fits all" way that goes beyond their intended scope and fails to take into account the context in which clinical psychology operates. We will then suggest that research credibility priorities in clinical psychology need to be situated within broader concerns about research waste. Pragmatically, the central aim of clinical psychological science is to develop effective means of preventing or treating mental health difficulties. Efficient research practices are those that optimise the quality, validity and speed, and minimise the cost (in every sense), of this treatment development pathway. While issues of rigour do need to be addressed in clinical psychology, the bigger problem is a failure to optimise this translational research pipeline. We will discuss how key areas of waste in the field are produced due to sub-optimal translation of findings from basic science into novel treatments (the t1 gap) and a failure to implement treatments shown to be effective in

controlled trials in real world settings (the t2 gap) (Cooksey, 2006). We will suggest ways to enhance the efficiency of this translational pipeline.

We take a UK perspective, identifying what we see as challenges for research credibility and efficiency in the context of the UK National Health Service and research funding bodies. We are aware that many but not all of the issues we raise will apply globally and should therefore be considered within each national context.

Emergence of the research credibility movement

It is increasingly recognised that many key findings in science generally, and psychology specifically, fail to replicate (e.g., Ionnidis, 2005; Chalmers & Glasizou, 2009; Open Science Collaboration; 2015; Camerer et al., 2018). Over 90% of respondents to a recent Nature survey agreed that there was a 'reproducibility crisis' across science (Baker, 2016). The endemic use of questionable research practices (QRPs) has also become apparent, which can lead to spurious false positive findings in basic psychological science (Simmons et al, 2011). At the design stage, these include chronic under-powering (Fraley & Vazire, 2014; Button et al., 2013) and inadequate protocol specification. At the data-collection stage, adaptive stopping rules are sometimes used (ceasing data collection when effects become significant). At the analysis stage, questionable practices include measure-hacking, p-hacking and selective exclusion (e.g., Ioannidis et al, 2014; Flake & Fried, submitted). At the write-up phase, problematic approaches include hypothesizing after the results are known, spin and providing insufficient detail to enable replication (Kerr, 1988; Glasizou et al., 2014). At the publication phase, there are 'file drawer' and citation bias problems (e.g., Ingre & Nilsonne, 2018). These QRPs until recently were seen as acceptable standard practice, were incentivized by the grant and publication reward structures, and most likely contributed to low rates of replicability in the field.

Clinical psychology research is not immune to these problems (e.g., Perepletchikova, Treat & Kazdin, 2008). Cuijpers and Cristea (2015) in a 'tongue in cheek' article discuss ways that may be (intentionally or unintentionally) used to prove a novel therapy is effective, even when it is not. These include: allegiance biases leading to enhanced supervision and training for the preferred therapy (Munder et al., 2013); maximising expectancy effects for the preferred treatment; exploiting weak spots in trial methodologies to favour the preferred treatment (e.g. selection of a 'straw-man' comparator, preferential allocation to treatment arm, non-blinded assessment, focusing on completer not intention-to-treat data, measure and p-hacking); and running small trials and only publishing the ones that work. Similar use of QRPs in clinical psychology and psychiatry research have been identified by other authors (Liechsenring et al., 2018; Reardon et al., 2018; DeVries et al., 2018; Turner et al., 2008).

In response to these concerns, a number of recommendations have been made about how to improve rigour and replicability in science (e.g., Munafo et al., 2017). These include: training in good methods to minimise QRPs; encouraging protocol pre-registration and/or registered reports; incentivising replication; ensuring studies are adequately powered; and strengthening the peer review process to ensure that it more rigorously detects bad practices (e.g., Munafo et al., 2017). Ways to adapt these recommendations to clinical psychological science have also been considered (Hopwood & Vazire, 2018; Tackett et al., 2018).

We wholeheartedly endorse these recommendations. However, we wish to raise some risks to innovation in science if these principles are over-rigidly applied, particularly to clinical fields where there is still significant uncertainty about how to operationalise and measure phenomena of interest. We begin by outlining what is required to maximise innovation in science.

Striking a balance between creativity and rigour in science

There is a long theoretical tradition considering how to optimise scientific development and innovation. This highlights a synergy (and sometimes inevitable tension) that exists between maximising creativity and preserving rigour (see reviews by Wagenmakers, Dutilh & Sarafoglou, 2018; Fiedler, 2018). Whewell (1840) proposed two mutually reinforcing modes of reasoning: inductive reasoning that generates a new creative leap, and deductive reasoning that then tests whether the leap was justified (i.e. the deduction verifies the induction). More recently, Kelly (1955a and 1955b) described two essential components of creativity: 'loosening' (breaking out of the shackles of existing knowledge and rules, developing unconventional ideas that go against the zeitgeist) and 'tightening' (critical evaluation, selection and then implementation of the ideas generated by 'loosening').

Optimal scientific development arguably involves striking a considered balance between (and appropriately sequencing) these inductive-loosening and deductive-tightening processes (Kaufman & Glaveanu, 2018). There must be space to engage in both induction and deduction, with induction tending to occur first to generate the input to the deductive process. When engaging in deductive, tightening processes, the optimal balance also needs to be struck between protecting from type I error (a false positive finding) and type II error (a false negative finding) and between ensuring internal validity (the experiment is free from error and any difference in measurement is due solely to the independent variable) and external validity (i.e. generalisability beyond experiment) of findings. As is well known, there is a trade-off between each of these (attempts to reduce type I error can enhance type II error; attempts to maximise internal validity can reduce external validity).

Having outlined the factors that maximise innovation in science, we will focus now evaluate the thesis that an overly rigid application of the research credibility movement could lead to some unanticipated costs. In particular, we will assess whether it could stifle

innovation and tilt the balance too far in the favour of protecting against type I error (and therefore maximise type II error). It is important to acknowledge we are predominantly focusing on risks rather than possible benefits of the research credibility movement in what follows. We are taking this 'one-sided' stance because others have already written eloquently about the positives associated with a thoughtful application of research rigour principals (e.g., Munafo et al., 2017; Hopwood & Vazire, 2018), while the negatives of an extreme application have rarely been explicitly evaluated. Our intention is to contribute to a dialectical synthesis around the merits of research credibility principles and to promote a balanced implementation in the field. In the following sections we will focus on depression intervention research to illustrate these risks.

The State of the Depression Field

Depression remains a chronic, recurrent, prevalent condition that is a major cause of disability despite extensive research efforts (Kessler et al., 2003; Ustun et al., 2004). Current psychological treatments for depression are only partially effective (with high rates of non-response or relapse after response; Cuijpers et al., 2008; Vittengl et al., 2007). It is not clear that current treatments are always superior to placebo-control when unpublished trials are taken into account (Cuijpers et al., 2014a, 2014b). There is uncertainty about whether therapies work by the distinct mechanisms their developers postulated (e.g. for cognitive-behavioural therapy, see Longmore & Worrell, 2007; Lorenzo-Luaces, German & DeRubeis, 2018) and there is debate about whether instead therapy 'non-specific' factors are driving improvement (Cuijpers, Reijnders & Huiber, 2018).

There has been a proliferation of different evidence-based treatments that appear to be equally effective, albeit with variation across treatments in the extent of their evidence-bases. Arguably, there have been no step-wise gains in the capacity to treat acute depression in the

past thirty years (Dunn & Roberts, 2016). There is an urgent need for treatment innovation and implementation.

It is unclear if this failure to differentiate treatments is due to genuine equivalence of efficacy, or to heterogeneity in depression. There is significant variability in the clinical presentation of depression, with approximately 1000 unique symptom combinations all leading to the same diagnosis (Fried & Nesse, 2016). Moreover, there is marked variation across the disease life span as a function of severity, chronicity, number of previous episodes, and previous treatment response ('staging' variables; Lorenzo-Lucaes, 2018). There are also distinct differences in patterns of comorbidity (Hasin et al., 2018). As a result, the adequacy of the current diagnostic system has been increasingly challenged (e.g., Insel et al., 2010; Fried et al., 2017; Borsboom, 2017; Hoffmann, Curtis & McNally, 2016) and difficulties in knowing how to best measure depression have been highlighted (Fried, 2017; Fried & Nesse, 2016; Fried et al., 2016). This heterogeneity is not only limited to client characteristics. There is also variability in how competently individual therapists deploy complex therapy protocols to treat depression (Saxon, Firth & Barkham, 2017) and how well these are implemented between different settings (Clark et al., 2018).

Given this heterogeneity, it is unsurprising that there has been a failure to find clear differences between treatments, or that treatments fail to be reliably superior to placebo. This is exacerbated in subsequent meta-analyses, which only partly control for heterogeneity between trials and do not take into account within-trial heterogeneity (for a balanced critique of meta-analyses, see Serghiou & Goodman, 2018). The results is that there is a danger in simply accepting the null in these trials. However, such a conclusion would ignore the possibility that different treatments are more effective for particular subtypes of clinical presentations, in particular contexts, or when delivered in targeted ways.

To move the depression field forwards, it is critical to gain a better understanding of boundary conditions between treatments, to help answer what works for whom and when. For example, an increasing body of work is now examining treatment selection in depression (for example, DeRubeis et al., 2014). Some of this work, using novel application of machine learning approaches, is beginning to reveal that even when trials show no difference between treatments in an overall sample, there can nevertheless be significant differences in treatment outcomes in particular subgroups (Lorenzo-Luaces et al., 2018; Cohen & Derubeis, 2018).

In many ways, the depression field, despite ancient historical roots, has many of the features of what Rozin (2001) calls a 'young' science. Rozin points out it is helpful for a young science to begin by identifying, conceptualising and learning how to measure underlying phenomena of interest. Ideally these are homogeneous and invariant (i.e. present in a similar way across individuals and contexts). These early phases are more descriptive and/or exploratory. Only once these steps have been achieved should there be a move to more 'mature' scientific methods that follow a confirmatory approach (using experimental methods to test clear hypotheses).

Depression research using deductive methods is vulnerable to the QRPs and replication issues identified by the research credibility movement (e.g., Turner et al., 2008; DeVries et al., 2018). However, the primary problem the depression treatment field faces is a lack of innovation of novel treatments that are then worthy of rigorous examination using these deductive methods. Moreover, when conducting deductive research there is a danger of a tilting the balance too much to protect against type I error rate (and therefore exacerbating type II error rate). We evaluate whether an overly rigid application of the recommendations of the research credibility movement could exacerbate these problems.

Potential risks of an extreme application of research credibility principles

Evaluating risks identified in the broader basic science literature

A recent symposium in *Perspectives in Psychological Science* debated the impact of research credibility principles on creativity in research and (alongside a number of likely benefits) identified a series of risks. 'Tightening' of research practices could result in: a reduction in the number of studies conducted because of increased workload required by new policies (Wai & Halpern, 2018); an over-emphasis on methods and analysis leading to an under-emphasis on theory construction (Fiedler, 2018); and researchers being increasingly reluctant to pursue 'risky' questions (Vazire, 2018). As a result there could be a shift towards 'little c creativity' (where scientists incrementally develop the current dominant theory rather than bring about paradigm shifts; Kuhn, 1962).

Of these concerns, we see a particular risk for clinical-psychological science with regards to prioritising theory construction that underpins subsequent treatment development. Popper (2005) articulated how initial theory development is often an inductive process that precedes formal scientific testing, but nevertheless is a rigorous process. In the early phases of research, scientists are guided by intuition or imagination, which they then subject to a process of rational criticism to begin to refine a theory iteratively. What is essential to this creative thinking is the coupling of imaginative freedom, a capacity to engage in highly critical thinking about the products of this imaginative freedom, and an intense interest in the problem (so the scientist is willing to immerse themselves in the issue and go through repeat iterations before they come up with the best account). What typically occurs is an iterative cycle between hypothesis, test and reflection (in relation to extant theory and literature), which ends when the optimal account is generated (i.e. the most consistent and powerfully explanatory model has been built). This is similar to the test-operate-test-exit (TOTE) cycle believed to guide behaviour in early cybernetic theory (Miller, Gallanter & Pribram, 1960). At times, each loop of this cycle can refine and simplify the problem space (Newell & Simon, 1972) in a linear fashion At other times, fresh information may be introduced into the system

that requires a new (and potentially non-linear) inductive leap to take place to be to account for it. This refinement of ideas prior to formal testing has been described by later writers as the 'pursuit' phase, where a theory-sketch has been generated but is not yet sufficiently well specified to be tested using formal scientific methods (see Chakrabarty, 2010). We need to ensure that this process of theory generation prior to empirical testing continues to be legitimised and scientists are given time and space to engage in it.

Relatedly, there is a risk that the field sees the only valid source of information for the development of theories that underpin psychological interventions as being basic science laboratory data. We are in danger of increasingly viewing therapy in purely scientific terms as value-free techniques, failing to take into account the social, economic, and political roots from which they emerged (Marks, 2017). For example, cognitive-behavioural perspectives were strongly influenced by stoicism – the idea that we are disturbed not by things but our view of things (Evans, 2012). The emergence of CBT also coincided with the cognitive revolution in other areas of science and the emergence of computing (Miller, 2003). Not all of these broader influences are easily reduced into testable experiments to run in the laboratory. We need to preserve a broad set of influences on our theorising, with experimental evidence being a key, but not the sole, source of theory development. Moreover, a theory should be evaluated not solely on the basis of how well it generates testable hypothesises that can be deductively examined in the laboratory but also the extent to which it and serves as a set of 'guiding principles' to inform clinical management in real-world settings. This echoes Lewin's views about the importance of ensuring a good theory is applicable (Lewin, 1943; 1951).

We will now consider a range of additional risks of an overly rigid application of research credibility principles.

Risk 1: Inhibiting Accidental Discovery

A risk of rigid implementation of the research credibility movement is that it may inhibit accidental discovery. The history of medicine is replete with examples of serendipitous discoveries revolutionizing practice (Ban, 2006). For example, Flemming first identified penicillin when a staphylococcus sample became contaminated and developed a mould culture that inhibited subsequent bacteria growth. The psychotropic effects of many drugs subsequently used to treat mental health were also discovered by accident, including lithium, tricyclic antidepressants, and monoamine oxidase inhibitors.

A strong form of this argument is that nearly all of the significant advances in our field were of this accidental form and very few emerged from a pure application of the deductive scientific method. We do not endorse this strong form. While accidental and free discovery should be enabled, it is also overstating the case to say that good discoveries only come about by these means. We do think that a theory driven 'experimental psychopathology' approach has delivered advances in clinical psychology (e.g. see Clark, 2004; Salkovskis, 2002). Moreover, even in the case of free or accidental discoveries, it is of course then essential to test them robustly.

Risk 2: Over-interpreting non-replications as true negative findings

A rigid application of research credibility principles could lead to overly simplistic thinking about how to interpret non-replications. The underlying assumption behind attempts to replicate is that there is a univariate, homogeneous phenomenon that can be measured and manipulated in a pure fashion. This may be possible in a basic biomedical science field. However, the complex psychological phenomena studied in clinical psychology are typically 'bounded' by person and context (Rozin, 2001), making direct replication (repeating the experiment in exactly the same way and seeing if the effects hold) virtually impossible. Instead, it is more realistic to aim for conceptual replications (varying aspects of the

experiment to see if the effects still hold; typically different settings or different subjects) (cf., Zwann, 2018).

Taking too narrow a view of replication in clinical psychology could lead to the dismissal of all 'near miss phenomena' (i.e. an interesting effect that is not significant in *a priori* planned analyses but is in some secondary ones) or results that do not fully replicate (i.e. effects that are found in one sample but not another). It is a logical error to conclude these 'near misses' are *always* 'true negatives' that are not worthy of further exploration. In particular, it may be that they are legitimate findings that are bounded by participant characteristics and context (and clarifying these boundary conditions is of value to the field). Of course, some findings in the present literature will be 'true negatives' that emerge from questionable practice and it would be a waste of resource to look for boundary effects of these phenomena. Balanced rather than extreme application of research credibility principles should be helpful in this regard, as over time the frequency of these 'true negatives' in the literature should reduce as QRPs are gradually eliminated and the field as a result will be able to have greater confidence in the findings that emerge in each individual study that is published.

Risk 3: Devaluing descriptive and exploratory analyses

An overly strict interpretation of research credibility recommendations regarding pre-registration could devalue descriptive and exploratory data analysis. It is already challenging to publish work of this kind in the field. For example, we rarely see in leading psychology journals work that clearly acknowledges its descriptive or exploratory origins and there is a pressure to provide a clean, narrative story that makes all analyses conducted seem as *a priori*, linear tests of a pre-determined nature (e.g., hypothesizing after the results are known [HARK-ing]; Kerr et al., 1983). If journals and reviewers apply simplistic binary decision rules to pre-registration (i.e. rejecting research that is not pre-registered,deviates in even

small ways from pre-registered protocols, or is not clearly deductive in nature) this could exacerbate the problem. A new 'file drawer' problem may emerge where descriptive and exploratory findings are never publishable.

This would be a pity, as in our experience descriptive and exploratory analyses can be an extremely productive phase of research, particularly for 'young' sciences in need of innovation (Rozin, 2001). A specific example of the value of descriptive approaches is provided by Skinner (1955) when writing about the trajectory of scientific discovery over his career. Core to his early work was careful description and characterisation of sometimes unexpected observations about the conditions in which in animal did or did not learn a behavioural contingency, without any need to be guided by or attempt to test a pre-specified learning theory (Skinner, 1949). Each of these observations was followed up rigorously but not solely by use of deductive hypothesis testing.

A criticism of exploratory analyses are that they are merely 'fishing expeditions'. However, at their best these analyses are often an iterative, theory informed discovery process that is both inductive and deductive in nature. The researcher observes an unexpected pattern of data, generates a revised or novel theory and set of hypotheses to account for this unexpected pattern, then conducts further analyses to test this theory. Through immersion in the data and becoming familiar with its constraints and possibilities, this can facilitate creativity and 'flow' that help to bring about new understandings to complex problems. As with initial theory generation, this typically follows a rapid TOTE cycle (Miller et al., 1960), which ends when the optimal explanation of the data are reached. If the analysis protocol had to be registered and new data prospectively collected to test it at each phase of this iterative discovery process, this would interrupt creative flow and significantly slow down evolution of that line of work.

Our hope and expectation, however, is that these risks regarding pre-registration will not materialise. Advocates of the credibility movement do clearly say that there is a place for exploratory, *post hoc* research as long as it is properly reported (Frankenhuis & Nettle, 2018). Moreover, there has been a move away from a binary distinction of research either being confirmatory or exploratory, to recognising there is a continuum. There is now increasing flexibility in the kind of pre-registration that can be completed (Nosek et al., 2018), for example allowing some data-driven decisions to be made (but encouraging people to articulate how these decisions will be made *a priori*) and also allowing pre-registration of secondary analysis protocols. Similarly, there is no reason not to encourage pre-registration of purely descriptive studies. If these more flexible kinds of pre-registration are widely adopted by journals and reviewers this may legitimise rather than inhibit well-conducted descriptive and exploratory research.

Risk 4: Rigid specification of clinical trial design

There is a risk that the way in which the field conducts clinical trials of interventions will become overly rigid and fail to take into account the particular requirements of mental health research (arguably akin to some of the problems associated with 'teaching to the test' in education; see Chomsky & Robichaud, 2014) . This could exacerbate type II error. In an elegant thought experiment Cuijpers and Cristea (2015) describe how one might go about designing a research programme to show a novel therapy to be effective when it is not (capitalising on type I error). We borrow from their framework here to consider the reverse: the features of a research programme designed to show a novel therapy to be no better than standard care, when it is in fact superior. We are intentionally taking a *reductio ad absurdum* stance here.

First, no one with any particular interest in a therapy should be allowed to evaluate it, ostensibly to eliminate allegiance bias but also ensuring people will not work hard to optimise

treatment delivery and trial data collection. Second, care should be taken not to provide extensive training and supervision for the novel therapy prior to starting the trial and instead this should be exactly matched against the treatment-as-usual arm (ignoring the fact that the therapists recruited into the treatment-as-usual arm have undergone extensive training in usual care prior to the trial and have then practised it clinically for a number of years). Third, to minimise measure-hacking, only a single outcome measure should be collected (despite the fact that we are really not sure what depression is, that multiple outcomes are valued by different stakeholders, and that most of our measures are flawed). Fourth, to minimise p-hacking and eliminate multiple comparisons, a rigid analysis protocol should be pre-specified with no room for data-driven modification (for example, if the primary measure is shown to have high rates of missingness, to have errors in its administration, or have poor reliability we should not move to another measure).

Fifth, any attempts to conduct moderation analyses to unpick what works for whom and when should be dismissed as a fishing expedition (despite the fact that there is overwhelming evidence of heterogeneity in the presentation of depression). Sixth, any variance between sites and therapists should be ignored as noise, not as key information to inform successful implementation of a complex intervention. Seventh, attempts to evaluate mechanism of action of complex multi-component treatments should either be avoided or be suitably reductionist, so that the burden to participants is minimised and so that no real opposition to the developers' preferred theory of change is inadvertently generated. An added advantage of this narrow focus of outcome, moderator and mediator measurement is that the trial data are of no use to any other researchers in the field for secondary analyses, so the principal investigator will not often be bothered by requests to share data.

Eighth, after the trial has been conducted, the outcomes should then be entered into trial-level meta-analyses. Meta-regression and sub-groups analyses should be run and

interpreted to support the null, without making reference to issues of chronic under-powering that exacerbate type II error. Individual patient data meta-analysis should be avoided, as it can throw up complexities in conclusions about whether to recommend a treatment or not at the population level.

Further, steps can be taken to maximise the chances that a treatment will fail to be implementable in real world practice, by focusing on maximising internal validity with no consideration of external validity and generalisation. An artificially 'clean' group of patients with no comorbidity and no differences in depression 'staging' should be recruited, ideally so tightly specified that these clients very rarely present in real-world practice. A very rigid treatment protocol should be specified with no scope for any tailoring to different presentations. There should be very prescriptive requirements about setting and supervision throughout the trial that cannot be easily recreated in real world settings.

Of course, no one in the research credibility movement, and we would hope nobody at all, is suggesting anything like the above would be a good way to conduct a trial in depression. However, there is a serious point here. The above is an extreme example of how to try and minimise type I error, without any consideration of type II error. Maximising internal validity of trials makes sense in areas of medicine where the phenomenon of interest is clearly operationalised, can be measured with a single outcome, the delivery of the intervention targeting the phenomena is simple and fixed, and implementation is not a challenge (e.g. see recommendations in Heneghan, Goldacre, & Mahtari, 2017). This is not the case in complex fields like mental health. This nuance may not always be appreciated by generic biomedical funding panels, who may expect mental health trials to follow the same ground rules as those used in 'cleaner', more mature areas of health research. Mental health trials may not be funded as a result.

We now make a number of recommendations to promote optimal trial design in mental health. The collection of multiple outcomes should be supported as a legitimate approach, ideally with a 'core set' agreed across trials. Indeed, triangulation of different measures of the same construct should be seen as good practice not 'data fishing' (e.g. Denzin, 1978). We agree with Guidi et al. (2018) who recommend multiple measurement of outcomes (clinimetrics) that move beyond simply indexing symptom relief to also cover social functioning, wellbeing, and patient satisfaction and cover staging variables. This broad array of outcomes beyond symptoms is often of particular importance to service-users (for examples in depression see: Zimmerman et al., 2006; Demyttenaere et al., 2015). Each additional outcome measure needs to add unique information to justify its inclusion (i.e. display incremental validity). Sensible suggestions have also been put forward by Flake and Fried (submitted) about selection of measures (defining the construct of interest, operationalising the construct, justifying the choice of measure, justifying any modifications to the measure, and justifying the creation of any bespoke measure) that will help preserve rigour.

Trials should have broad recruitment criteria that reflect real-world presentations of mental health problems. Given the chronic and/or recurrent nature of depression, there needs to be multiple outcome measurement over a long time period (rather than rigidly pre-specifying one time point as the primary outcome and not examining longer term benefits). A good example of this is follow-up analyses of the COBALT trial, which demonstrated five-year clinical and cost-effectiveness of CBT (compared to treatment as usual) for cases of depression that had not previously responded to medication (Wiles et al., 2016).

A particularly important area for progress in the field is to ensure trials are optimally designed to explore moderators and mediators. It may be the case that no single treatment will ever end up being superior for all cases of depression given its inherent heterogeneity. It

is therefore of value to be able to better match the right treatment to the right presentation. A wide-ranging set of individual moderators should be collected (and ideally standardised across trials) to help establish what works for whom and when. Given clear evidence that staging (severity, chronicity, number of episodes) impacts on treatment response, the list of moderators should include staging variables (Lorenzo-Luaces, 2018; Guidi et al., 2018). Given increasing evidence from network psychopathology that not all symptoms of disorders are 'created equal' (e.g., Fried et al., 2017), whether particular elevations in particular symptom clusters predicts outcome should be examined. It is also important to capture contextual moderators (including therapist and service-context), which can be achieved by using methods from realist complex intervention science (Fletcher et al., 2016) and also ensuring a robust process evaluation is built into trial design (see Moore et al., 2014).

There should be an emphasis on the development of ways of analysing moderation to maximise the capacity to match the right treatment to the right patient (Cohen & DeRubeis, 2018). Robust validation of putative mediators and ways to evaluate them should occur using laboratory methods and only then be included in trials when they are well specified. Mediators and outcomes should be assessed at multiple points during treatment, given the importance of establishing temporal precedence and the difficulty in predicting *a priori* exactly when in treatment change will occur. Analysis methods should be developed that make it possible to allow for individual differences in when this change comes about, rather than assuming this is fixed for all individuals (for general recommendations around mediation analysis, see Emsley et al., 2010; Kraemer et al., 2002; Hayes & Rockwood, 2017). Trials should be adequately powered to conduct these mediation and moderation analyses (and ideally to examine site and therapist effects; Spirito et al., 2009; Kraemer & Robinson, 2005; Magnusson, Andersson, & Carlbring, 2018), although we recognise this may not be pragmatically possible in all cases. One approach might be to form large, cross-national

consortia to enable trials with sufficient sample sizes to be run that are powered to explore these issues. Such an approach is being successfully followed in other areas of psychiatric research (for example, the Psychiatric Genomics Consortium: see https://www.med.unc.edu/pgc/about-us/).

Meta-analytic techniques that are sensitive for examining moderation at the individual participant rather than trial level should be utilized where possible. For example, using individual patient data meta-analysis, it has been demonstrated that Mindfulness Based Cognitive Therapy is more effective as a relapse prevention treatment for depression (relative to maintenance anti-depressant medication) in participants with more marked residual depression symptoms (Kuyken et al., 2016). Further, another individual patient data meta-analysis in older adults found a superiority of anti-depressants to placebo only in those with a long illness duration and at least moderate depression severity (Nelson, Delucchi & Schneider, 2013). At present, individual patient data meta-analyses are hampered by the extensive variation in trial design and measurement, so establishing a standard set of moderator measures to include across different future trials will be beneficial.

Careful consultation with patients should occur at the trial design phase to find ways to ensure measurement burden is not overly onerous and to maximise chances of high rates of data completeness (e.g., see Bodart et al, 2018). Given that this will create multi-factorial data sets, care should be taken at the protocol registration stage to be clear about what are primary and outcome measures (or what decisions rules will be used to inform selection of primary outcome after data collection is completed) and how analyses will be conducted.

The correct control group should be selected based on the question of interest and how well developed the field is (see Gold et al., 2017), not just automatically defaulting to a two-arm comparison of two active therapies in all scenarios. Where the key question is to understand how and for whom a treatment works, it may be appropriate to use an

'experimental' control condition (for example, one that is identical to the treatment of interest apart from some particular processes/elements of interest). Where it is uncertain how effective standard care is relative to no intervention, it may be appropriate to include a third arm like a waitlist control (although for ethical reasons it is inappropriate to deny clients at risk from treatments that we do definitely know are effective; Gold et al., 2017). Where there is no current standard of care (for example, evaluating what to do with treatment resistant clients after all standard treatments have been exhausted), it may be appropriate to have a treatment as usual or no intervention control condition.

Pilot trials are often conducted prior to a definitive trial to determine if it is feasible to run. While the recommendation is not to analyse clinical outcomes in pilot trials as they are under-powered (Thabane et al., 2010), there may nevertheless be other ways to evaluate proof-of-concept of the intervention including interpreting confidence intervals and using Bayesian methods (Lee et al., 2014). There may also be merit in analysing pilot trial data at the individual participant level, for example assessing reliable and clinically significant improvement and deterioration rates (cf. Jacobson & Truax, 1991). If data are collected at multiple intervals (for example, weekly sessions), with the addition of a baseline phase it may also be possible to conduct intensive time series analyses at the individual level.

It should be accepted as legitimate practice for intervention developers to be involved in evaluating them, at least in the early stages of the development pipeline. However, there is no way to escape the fact that treatment developers are likely to have direct and indirect financial and career gains if their treatment is widely adopted, meaning they are vulnerable to intentional or unintentional bias. When developers do evaluate their own treatments, they should clearly document steps they have taken to minimise allegiance bias and ensure equipoise between arms. For example, adversarial collaboration may be useful (where primary researchers who each favour different treatments jointly work on the same trial, with

both centres running both arms) (see Leykin & DeRubeis, 2009; Haaga, 2009). For a treatment to be considered well validated, while a portion of trials can be conducted by the treatment developers, some should be conducted by genuinely neutral research groups.

When funding panels review mental health trials, a specialist in this area should be part of the committee and also panel members from broader bio-medical fields should be made aware of the 'messy' context in which mental health operates and why this changes the requirements of good trial design.

Having now reviewed the extent to which an overly-rigid focus upon reducing type I error might exacerbate type II error, we will now extend the focus to look at broader issues of research waste in the clinical psychology field.

Part 2: Broadening the focus to other areas of research waste

Psychology to date has primarily concerned itself with problems of non-replication and associated QRPs that may underpin these. However, this is only one part of a broader set of factors contributing to research waste (see Chalmers et al., 2014; Al-Shahi et al., 2014; Chan et al., 2014). By shining the light solely on the replication issue in psychology, there is a risk that other areas of waste will be left in the dark (and therefore will not be minimised). We will now review other sources of waste of relevant to the development of psychological therapies.

We see therapies as optimally emerging via a translational research pathway, with basic research informing treatment development, evaluation and then implementation. It is currently recognised that a disappointingly small amount of basic science research goes on to inform healthcare practice in this way (Grimshaw et al., 2012) and that it takes too long to move through this pathway (Morris et al., 2011), meaning treatments may be obsolete or no longer fit for context at the end of the process. Anything that enhances the efficiency of this

pipeline (and eliminates waste) will therefore be of benefit. Two key rate limiting steps identified in the treatment development pipeline are translating basic science insight into new treatments protocols (the t1 gap; 'the valley of death') and then ensuring that novel treatment protocols, shown to work in ideal settings, are implemented in a real world context (the t2 gap) (Cooksey, 1996; Butler, 2008; Coller & Califf, 2009). We will review ways to overcome these gaps. In addition, we will discuss a third issue regarding how to allow organic growth of psychological therapies after the development pipeline has been completed.

These other sources of research waste are best seen as intimately intertwined with the research credibility movement. A thoughtful, flexible application of research credibility principles will help, while an overly rigid implementation of research credibility principles will hinder, progress in reducing research waste in each of these areas. While these issues are interlinked, what counts as optimal methodological rigour may look somewhat different at each phase of the translational pipeline (development, evaluation and implementation).

Source of waste 1: Not optimising treatment development

Psychological therapies are complex interventions, which are not straightforward to develop. It is a mistake to neglect this development phase. The MRC complex intervention framework (Craig et al., 2006) is explicit that excessive focus upon definitive trial evaluation, neglecting the development, piloting, and implementation phases, can result in interventions that are: less likely to be effective; harder to evaluate; less likely to be implemented; and less likely to be worth implementing in the first place. In other words, not optimising a treatment prior to trial significantly inflates the risk it will not be any better than existing treatments.

Clark (2004) and Salkovskis (2002) both write eloquently about the systematic steps they followed to develop novel treatments for anxiety disorders, via a creative synergy between the clinic and the laboratory. There is a rapid, iterative cycle between theory

development, experimental science and treatment development, at the core of which is an attention to phenomenology when working with patients.

Therefore, the optimal context in which to develop new therapies is to have time and space to foster a creative synergy between clinical practice, research, and broader contextual influences. In modern day clinical-academia, this is hard to achieve for a variety of reasons. First, it is challenging to gain funding to develop novel treatments – funders tend to cover either the basic science arena or the treatment evaluation arena. Asa result, it is difficult to do the iterative development work to translate basic science findings into a novel treatment to evaluate. Second, it is difficult (in a UK context at least) to have a split academic and clinical post. After clinical training individuals are typically forced to choose either a pure research or clinical pathway. This mean researchers lose their links to 'grass roots' practice, both inhibiting their creativity and inflating the risk they will develop a treatment that is not fit for real world context. The UK National Institute of Health Research (NIHR) has recognized this as a significant problem for allied health professionals (NIHR trainee coordinating centre, 2017). Third, if an individual has found a way to develop a novel treatment with encouraging preliminary data, it can then be challenging to publish these findings. Journals often view this development work as insufficiently robust, so reject it at the review phase.

These factors may impact on the career choices that researchers make. Scientists may be put off becoming treatment developers at all, instead choosing to operate in the basic science arena or the trials arena. Alternatively, researchers may rush the development of an intervention and take this to RCT evaluation before it is optimised (both in terms of how it works and how to implement it effectively). This then increases the odds of finding a null result in the subsequent trial (even if that trial is conducted in a way that is fully compliant with research credibility guidelines).

To minimise this source of waste, greater emphasis should be placed on systematic treatment development. The MRC guidelines for complex intervention development can be a helpful starting point to follow here (Craig et al., 2006). These recommend: being clear what the treatment is targeting; building a coherent underlying theoretical basis to the treatment; having a clear, well specified protocol that others can follow; and being clear how change will be measured. Further, any uncertainties about whether the planned definitive trial evaluation can be conducted (including recruitment, retention, intervention acceptability and feasibility, likely effect size and variability) should be resolved in piloting and feasibility work before proceeding to evaluation. These guidelines are of optimal utility when they are seen as a useful set of guiding principles which can be adapted to each specific intervention context, rather than a prescriptive set of rules that must be followed in a 'tick-box' fashion.

Moreover, the MRC guidelines are arguably under-specified as to how complex protocols are optimally designed. The steps outlined by Clark (2004) can be particularly helpful when developing psychological therapies: i) use clinical interviews and experimental paradigms to identify core mechanisms triggering and maintaining a disorder; ii) construct theoretical accounts as to why these mechanisms do not self-correct; iii) use experimental studies to test these hypothesized maintaining factors; and iv) develop specialised treatments that reverse these maintaining factors (sometimes by direct translation of experimental manipulation procedures used in the earlier step). We would add to this the need to test evidence for an underlying mechanism by establishing it is cross-sectionally and prospectively linked to the target outcome using questionnaire and experience-sampling designs, and showing that manipulating it in the laboratory and in real world settings changes the target outcome (Dunn et al., 2017). Well conducted basic science that thoughtfully rather than rigidly follows research credibility principles is likely to be helpful in this regard, as it will help ensure that any mechanisms that treatments target are more likely to have a sound

evidence-based footing. The notion of 'full-cycle' evaluation from social psychology may optimise the utility of such work (e.g. Mortensen & Cialdini, 2010). This recommends that researchers start with naturalistic observation to establish the present of an effect in the real world. They then develop a theory to determine what processes underlie this effect and use experimental methods to verify this theory. Critically, they then return to observational work in the natural environment to corroborate the experimental findings.

Case series methods can be particularly helpful in the preliminary evaluation of treatments, being ideally suited to assess treatment acceptability and feasibility and provide preliminary evidence of effectiveness and proof-of-concept without requiring significant investment of resources (Kazdin, 2011; Morley, 2017). Use of a randomised multiple baseline design (randomising individuals to different lengths of baseline phase before starting treatment) helps to differentiate between a genuine treatment effect and natural recovery over time or other confounding factors (Kratochwill & Levin, 2010). Intensive time series analyses have adequate power to test statistically the efficacy of an intervention for an individual participant (Borckardt et al., 2008). By replicating findings across a series of individual cases, this begins to assess whether findings are generalizable (for example using single-case meta-analytic techniques; Jamshidi et al., 2018). There is potential for multi-centre collaboration here, with the same interventions being evaluated in different settings to explore context effects and to 'road test' implementation issues. It is possible to refine the intervention between case series waves, allowing for rapid optimisation of treatment protocols.

Case series can also help identify treatment non-responders, who can be overlooked within the overall effect of a group-based design. Case series can be further enhanced by incorporating qualitative methods, allowing for detailed exploration of patient views on feasibility, acceptability, efficacy, mechanisms of action, reasons for non-response, and ways in which the treatment can be improved (Onghena, Maes & Heyvaert, 2018). It is also

possible to estimate pre to post effect sizes at the group level in these case series, which ideally should be at least of a large magnitude according to Cohen's rules of thumb (Cohen, 1988).

Case series continuation rules can be pre-specified (e.g., the treatment is acceptable to patients and a majority will complete a minimum acceptable course; at least 50% of participants show reliable and clinically significant change; on average a large effect size is observed; therapists can be trained to deliver with minimum adequate fidelity) to determine whether a protocol needs further refinement or is ready to proceed to RCT evaluation. If required, further iterations of the protocol can be made and the case series can be repeated on the optimised protocol before proceeding to the trial stage. Arguably a definitive trial should be conducted only where there is little uncertainty that the treatment will be effective. The purpose of a definitive trial should be to establish how the treatment performs relative to other treatments (Clark, 2004).

Funders should ensure they have 'balanced portfolios of investment' that fall in different places on the translational pipeline (critically including treatment development and not just focusing on treatment evaluation). There should be more explicit consideration of value for money in the work funded (for example, adaptation of value of information analysis methods from health economics; Tuffaha, Gordon & Scuffham, 2014). Wherever possible, joint clinical-academic posts should be supported to allow treatment developers to be clinically active whilst they develop new treatments. In a UK context, this would be best served by joint posts between universities and the NHS, similar to those available to medical doctors.

Journal editors and reviewers should be encouraged to publish development work where it is done to adequate standards and reported in an appropriately conservative fashion given its preliminary nature. For example, journals should be willing to consider rigorously

conducted case series using intensive time series analyses that follow published methodological guidelines; thoughtfully analysed qualitative studies of stakeholder views of the problem; and pilot/feasibility studies that establish the planned definitive trial can be conducted and which include appropriate proof of concept analyses given sample size limitations.

Source of waste 2: Neglecting implementation

There is an acknowledged divide between clinical work in an 'ivory tower' academic setting and at the clinical 'coal face' (Stirman, Gutner, Langdon & Grahm, 2016). The argument is that treatments are developed in the ivory tower that can only work with a carefully screened, homogeneous group of patients where therapy is delivered under ideal conditions. These treatments are not always implementable in real world settings, and if they can be implemented they are significantly less effective. This may reflect two distinct underlying issues: i) the treatments designed in the 'ivory tower' may not actually be fit for the real world service context as the developers did not consult widely enough at the design stage; ii) the means by which we are disseminating the treatment into the real world context are not optimised (therapist selection, training and supervision; service context). Each of these issues raises slightly different challenges and responses.

With regards to how to design an intervention that is fit for the real world context, it is important to consider implementation from the outset rather than view it as something that is thought about only after a trial has shown that a treatment works in ideal settings. This can be facilitated by ensuring key stakeholders such as patients, clinicians and commissioners are consulted at all stages (Dunn, 2017). In the UK, the NIHR INVOLVE programme has helped ensure that patient consultation is now embedded into all stages of the research process in NIHR funded projects (INVOLVE, 2012). There are emerging examples of intervention co-design in the mental health field, where patients and researchers work together to develop the

treatment (e.g. see Nakarada-Kordic et al., 2017; Larkin, Boden & Newton, 2015).While the benefits of patient involvement are only starting to be empirically evaluated, and a sense of how best to use PPI input is still evolving, results are encouraging (see Brett et al., 2014)

There is a range of formal frameworks that can guide this initial design process. For example, the intervention mapping approach – a framework pioneered in health psychology to aid the effective development and roll out of public health interventions – emphasises the importance of this co-design process (Bartholomew et al., 2016). It can be particularly useful to have multiple stakeholder perspectives to help resolve conflicts of interest.

Likely affordability and cost-effectiveness of an intervention should be considered from the outset. There is little point in developing a novel therapy that is too expensive to implement in its target context and/or that is worse value for money than existing treatments. Therefore, health economic methods should be incorporated from an early stage, including detailed costing of intervention delivery and development of health economic models to look at likely long term cost-effectiveness. Ideally these models should take a broad societal perspective and consider likely savings over the longer term, given the widespread impacts of mental health and the chronic, relapsing course of presentations. Arguably, only treatments that emerge as likely to fall under acceptable absolute cost thresholds and to be at least neutral in terms of cost-effectiveness relative to standard care should be allowed to proceed to definitive trial phase.

With regards to how to disseminate an appropriate treatment into the real-world context after it has been trialled, implementation science perspectives may be helpful. Implementation studies focus on the rates and quality of use of evidence-based approaches, rather than whether those evidence based approaches are clinically effective in their own right (Bauer et al., 2015). Methods used in such implementation research can include both process evaluation (evaluating implementation without any direct intention to change the

ongoing process) and formative evaluation (evaluating implementation and feeding back to the implementation team to try and enhance ongoing practice) (Bauer et al., 2015). If it is not possible to conduct formal implementation controlled trials in a real-world settings, interrupted time series designs may be appropriate (measurement of outcome occurs at multiple time points before and after the implementation effort) (Bauer et al., 2015). One example of a helpful framework derived from implementation science is Normalisation Process Theory (May, 2013), which discusses ways to ensure that a health care intervention becomes a routine part of clinical practice over the long term. In particular, it is useful to think of the sense-making people do when first tasked with implementing a new practice, the relational work that is necessary to build and sustain a community of practice around a complex intervention, the operational work that helps individuals enact a set of practices, and the appraisal work that individuals do when evaluating the impacts of a set of practices on themselves and others. This methodology has been increasingly applied in the complex interventions field to aid intervention development and implementation (May et al., 2018).

In addition, there needs to be further research into the best methods to select, train and supervise effective therapists for a given treatment approach, how to ensure ongoing assessment and monitoring of therapist competence, and what counts as optimal dosage for treatment (e.g., see Shafran et al., 2009). Analysis of routine registry data after a treatment has been rolled out can provide a 'natural experiment' to identify what factors predict successful implementation (for example, looking at differences in outcomes between therapists and site and seeing what underpins these differences; e.g., Clark et al., 2018). As this relies on large scale roll out of therapy provision, it may be viable only in contexts with well-developed public mental health systems that capture routine registry data.

Source of waste 3: Inhibiting organic evolution of treatment

A neglected issue is how to allow complex interventions to evolve after the definitive trials evaluating them have been completed. The delivery of most drug and biomedical procedures are relatively fixed. In contrast, psychological therapies often continue to be refined in the years after the definitive trials have been conducted, informed by clinicians' experiences of implementation and through integration of other non-trial sources of empirical evidence. For example, CBT for depression has been enriched over the years by mechanistic insights in basic science being translated into novel interventions (e.g. rumination being identified as a key maintenance factor; Watkins, 2008). Strictly speaking, integrating these techniques into routine CBT practice means practice is no longer truly evidence-based, as the protocols have not been trialled with these novel elements included. However, these adaptations are clearly evidence informed. It would be counter-productive to prevent these techniques from being used.

What tends to happen is that when innovation of the above kind happens, researchers are nudged into labelling each of these adaptations as a distinct new therapy, as this is the only way to gain funding to develop and evaluate them properly. They go on to conduct new trials, for example head-to-head evaluations to see if the new treatment is clearly better than what came before. This process is resource intensive and slow. Given the incremental rather than stepwise nature of these refinements in some cases, it is unlikely that anything other than small effect size differences will be found (and very large and costly trials would be needed to demonstrate this). This proliferation of multiple protocols may not be helpful for the field, not least because clinicians then end up having to learn and choose between multiple protocols (when it would be more efficient to train them in a universal way of working that could be tailored based on client presentation).

Rapid, resource-efficient ways of refining existing treatments like CBT need to be developed that avoid this artificial proliferation of 'distinct' protocols and allow existing

protocols to continue to develop rather than calcify. It is not viable to go through the conventional trial pipeline for each new adaptation that is made, comparing them to the standard protocol. It is likely that each adaptation will result in a 'marginal gain' and probably only for the subset of clients where that particular mechanism/feature is present, meaning head-to-head comparisons are unlikely to show superiority. We are unsure of the optimal way to achieve this goal, and here make some tentative suggestions. One approach could be stepped wedge 'training' trial designs in real world settings (Hemming, Haines, Chilton, Girling & Lilford, 2014). For example, CBT clinicians (such as high intensity Improving Access to Psychological Therapies [IAPT] workforces) could be trained in novel additional techniques to add to the standard depression protocols at the cluster level, and the outcomes be evaluated. Another approach could be to conduct single session experimental designs, for example adding a novel session into existing treatment and carefully monitoring the impact this has on outcomes (see Clark, 2004). Further, there may be value in adaptive rolling designs, where multiple treatment options are tested simultaneously and sequential Bayesian analysis removes poorly performing arms (see Blackwell, Woud, Margraf & Schonbrodt, 2018). However, such an approach may not be easily operable except in the e-health domain. A potentially more radical solution is to move to a 'process-based' approach to evaluating therapies. The emphasis is on identifying the underlying mechanisms that maintain distress and what therapeutic procedures are effective at altering these mechanisms, irrespective of what school of therapy is being practised (see Hoffmann & Hayes, 2019).

Additional sources of waste

There are a range of other sources of waste that are not specific to clinical psychology, so we only briefly allude to them here. There is emerging recognition that the current grant system may not optimise efficiency: the efforts researchers make in writing unsuccessful proposals more than offset the gains made from selecting the best proposals,

especially when only a few proposals can be funded (Gross & Berstrom, 2019). There is now interest in alternative funding methods like partial lotteries or funding researchers based on past scientific success, each with different pros and cons (Gross & Bergstrom, 2019; Smaldino et al., 2019). While we have not seen anything written on the topic, there may be significant waste in publication practices also, for example, time wasted writing multiple submissions of the same piece to different journals, working down the journal impact factor hierarchy to ensure papers are published in the most career-enhancing outlets. Arguably it is best practice to publish more rapidly. There are also sources of waste in the regulation of science, for example inefficient administrative process around gaining ethics approval, negotiating intellectual property, and difficulties in optimising use of routine registry data due to failure to implement electronic records, to name but a few (see Al-Shahi Salman et al., 2014).

Conclusion

In this article we have discussed the evolution of the research credibility movement, which we agree is very helpful in eliminating type I error in fields characterised by high degrees of theoretical innovation but limited rigorous evaluation of these ideas. However, we have argued there are risks to inhibiting innovation and creativity in clinical psychology if these principles are applied too rigidly by funders, journals or scientists. These include inhibiting accidental discovery; placing treatment and theory development in a scientific vacuum and neglecting broader cultural influences; over interpreting non-replications as true negative findings; devaluing exploratory analyses; and conducting rigid clinical trials that follow a biomedical ideal and fail to take into account the heterogeneity of mental health. We have also reviewed broader contributors to research waste when developing

psychological therapies. These include not investing in initial treatment development, not optimising treatment implementation, and not allowing treatments to grow organically.

The net product of all these factors could be a significant reduction in the extent and the pace of innovation that occurs in the clinical field. Fewer novel treatments will emerge and those that do could take longer to move through the translational pipeline. Our hope is that by discussing these issues the benefits of the research credibility movement can be gained without leading to inadvertent costs. Moreover, we hope that the broader translational pipeline can be optimised. While we have largely restricted our discussion to depression, it may be that many of these concerns would be relevant for other areas of the clinical psychology field also characterised by significant heterogeneity and a need for innovation.

Acknowledgements

Dunn, Wright, O'Mahen and Brown met for a day to discuss the themes and issues to address in this article. Dunn wrote the initial manuscript and Wright, O'Mahen and Brown commented on and edited multiple drafts. The authors received no funding from an external source and declare no conflict of interest. Thanks to Sophie Dunn for proof reading.

References

Al-Shahi Salman R, Beller E, Kagan J, et al….. & Chalmers, I. (2014) Increasing value and reducing waste in biomedical research regulation and management. *Lancet, 383,* 176-185.

Baker, M. (2015) Is there a reproducibility crisis. A Nature survey lifts the lid on how researchers view the 'crisis'. *Nature, 533,* 452-454

Ban, T. A. (2006). The role of serendipity in drug discovery. *Dialogues in Clinical Neuroscience, 8*, 335-344.

Bartholomew, E. L. K., Markham, C. M., Ruiter, R. A. C., Fernàndez, M.E., Kok, G. &, Parcel, G. S. (2016) *Planning health promotion programs: An Intervention Mapping approach*. Hoboken, NJ: Wiley.

Bauer, M. S., Damschroder, L., Hagedorn, H., Smith, J., Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology, 3*, 32.

Blackwell, S. E., Woud, M., Margraf, J., & Schönbrodt, F. D. (2018). Introducing the leapfrog design: A simple Bayesian adaptive rolling trial design for accelerated treatment development and optimization. https://doi.org/10.31234/osf.io/zywpr

Bodart, S., Byrom, B., Crescioni, M., Eremenco, S., Flood, E., ePRO consortium (2018). Perceived burden of completion of patient-reported outcome measures in clinical trials: Results of a preliminary study. *Therapeutic Innovation & Regulatory Science,* DOI: 10.1177/2168479018788053

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63,* 77.

Borsboom, D. (2017) A network theory of mental disorders. *World Psychiatry, 16*, 5-13.

Brett, J., Staniszewska, S., Mockford, C., Herron-Marx, S., Hughes, J., Tysall, C., & Suleman, T. (2014). Mapping the impact of patient and public involvement on health and social care research: a systematic review. *Health Expectations, 17*, 637-650.

Butler, D. (2008)Translational research: crossing the valley of death . *Nature, 453*, 840–842 .

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., Munafo, M. R. (2013). Power failures: why small sample size undermines the reliability of neuroscience. *Nature Reviews, 14*, 365-376.

Byford, S., & Raftery, J. (1998). Perspectives in economic evaluation. *British Medical Journal*, *316*, 1529-1530

Camerer, C. F., Dreber, A., Holzmeister, F., et al….. & Wu, H. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637-644.

Chakrabarty, M. (2010) Karl Popper on theory creation. *International Journal of Arts and Sciences, 3*, 377-392.

Chalmers I, Glasziou P. (2009) Avoidable waste in the production and reporting of research evidence. *Lancet, 374*, 86–89.

Chalmers, I., Bracken, M .B., Djulbegovic, B., et al….. & Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *Lancet, 383*, 156-165.

Chomsky, N., & Robichaud, A. (2014). Standardized Testing as an Assault on Humanism and Critical Thinking in Education. *Radical Pedagogy, 11*, 3-11.

Clark, D. M. (2004). Developing new treatments: On the interplay between theories, experimental science and clinical innovation. *Behaviour Research & Therapy, 42,* 1089-1104.

Clark, D. M., Canvin, L., Green, J., Laard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *The Lancet, 391,* 679-676.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112,* 155-159.

Cohen, Z. D., & DeRubeis, R. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology, 14,* 209-236.

Coller, B. S., & Califf R. M. (2009) Traversing the valley of death: a guide to assessing prospects for translational success. *Science Translational Medicine,* 1, 1–5.

Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. (2008)Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal, 337,* a1655.

Cooksey, D. (2006) A review of UK health research funding.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology, 76,* 909-922.

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014a). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *Journal of Affective Disorders, 159,* 118-126.

Cuijpers, P., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014b). Comparison of psychotherapies for adult depression to pill placebo control groups: a meta-analysis. *Psychological Medicine, 44*, 685-695.

Cuijpers, P., & Cristea, I. A. (2016). How to prove that your therapy is effective, even when it is not: a guideline. *Epidemiology and Psychiatric Science, 25*, 428-435.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2013). The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLOS One, 9*, e83875.

Demyttenaere, K., Donneau, A. F., Albert, A., Ansseau, M., Constant, E., & Van Heeringen, K. (2015). What is important in being cured from depression? Discordance between physicians and patients (1). *Journal of Affective Disorders, 174*, 390-396.

Denzin, N. K. (1978) *The Research Act* (2nd edition). New York: McGraw-Hill

De Vries, Y. A., Roest, A. M., de Jonge, P., Cuijpers, P., Munafo, M. R., & Bastiaansen, J. A. 2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine, 48*, 2453-2455

Dunn, B. D. (2012). Helping depressed clients reconnect to positive emotion experience: current insights and future directions. *Clinical Psychology & Psychotherapy*, *19*, 326-340.

Dunn B. D., Roberts H (2016). Improving the capacity to treat depression using talking therapies: Setting a positive clinical psychology agenda in Wood A, Johnson J (eds.) *Handbook of Positive Clinical Psychology.*

Dunn, B. D. (2017) Opportunities and challenges for the emerging field of positive emotion regulation: a commentary on the special edition on positive emotions and cognitions in clinical psychology. *Cognitive Therapy and Research, 41*, 469-478.

Dunn, B. D. (in press) Augmenting Cognitive Behavioural Therapy to Target the Anhedonic Symptoms of Depression. In J. Gruber (Eds.) Oxford Handbook of Positive Emotion and Psychopathology.

Emsley, R., Dunn, G., & White, I. R. (200x). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. Statistical Methods in Medical Research, 19, 237-230.

Evans J. (2012) *Philosophy for Life and Other Dangerous Situations.* London: Rider Books.

Fiedler, K. (2018) The creative cycle and the growth of psychological science. *Perspectives on Psychological Science, 13*, 433-438.

Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. https://doi.org/10.31234/osf.io/hs7wm

Fraley, R. C., & Vazire, S. (2014). The N-Pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PlosOne, 9*, 10, e109019.

Fletcher, A., Jamal, F., Moore, G., Evans, R. E., Murphy, S., & Bonell, C. (2016). Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. *Evaluation, 22*, 286-303.

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology, 51,* 1-10.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191-197.

Fried, E. I., Nesse, R. M. (2016). Depression sum-scores don't add up: why analysing specific depression symptoms is essential. *BMC Medicine, 13*:72.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, Borsboom, D. (2016). Measuring depression over time …. Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28,* 1354-1367

Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science,* https://doi.org/10.1177%2F1745691618767878

Glasziou P, Altman DG, Bossuyt P., et al……..& Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *Lancet, 383,* 267-276.

Gold, S., Enck, P., Hasselmann, H., Friede, T., Ulrich, U. H., Mohr, D. C. (2017). Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. *Lancet Psychiatry, 4,* 725-732.

Grimshaw, J. M., Eccles, M. P., Lavis, J. N., Hill, S. J., & Squires, E. J. (2012). Knowledge translation of research findings. *Implementation Science, 7,* 50.

Gross, K. & Bergstrom, C. T. (2019). Contest model highlight inherent inefficiencies of scientific funding competitions. *Plos Biology*. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000065

Guidi, J., Brakemeier, B., Bockting, C. L. H., et al….. & Fava, G. A. (2018). Methodological recommendations for tirals of psychological interventions. *Psychotherapy and Psychosomatics, 87,* 276-284.

Haaga, D. A. F. (2009) A timely reminder on research design and interpretation. *Clinical Psychology Science and Practice, 16,* 66-68.

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

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., Grant, B. F. (2018). Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the united states. *JAMA Psychiatry, 75,* 336-346.

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis and reporting. *British Medical Journal, 350,* h391.

Heneghan, C., Goldacre, B., & Mahtari, K. R. (2017). Why clinical trial outcomes fail to translate into benefits for patients. *Trials, 18,* 122.

Hofmann, S. G., Curtiss, J., & McNally, R. J. (2016). A complex network perspective on clinical science. *Perspectives in Psychological Science, 11,* 597-605.

Hoffman, S. G., & Hayes, S. C. (2019). The future of intervention science: Process-based therapy. *Clinical Psychological Science, 7,* 37-50.

Hopwood, C. J., & Vazire, S. (2018). Reproducibility in clinical psychology. In Wright, A. G. C., and Hallquist, M. N. (Eds) *Handbook of Research Methods in Clinical Psychology*.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Ingre, M., & Nilsonne, G. (2018). Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate. *Royal Society Open Science, 5*, 181190.

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., et al……. & Tibshirani, R. (2014) Increasing value and reducing waste in research design, conduct, and analysis. *Lancet, 383*, 166-175.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C. & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry, 167*, 748-751.

INVOLVE. (2012) *Briefing notes for researchers: involving the public in NHS, public health and social care research.*

Jamishidi, L., Heyvaert, M., Declercq, L., et al……. & Van den Northgate. Methodological quality of meta-analyses of single-case experimental designs. *Research in Developmental Disabilities, 79*, 97-115.

Kaufman, J. C., & Glaveanu, V. P. (2018) The road to uncreative science is paved with good intentions: Ideas, implementations, and uneasy balances. *Perspectives on Psychological Science, 13*, 457-465.

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

Kelly, G. A. (1955a). *The Psychology of Personal Constructs: Vol. 1. Theory and Personality*. Oxford, England: Norton.

Kelly, G. A. (1955b). *The Psychology of Personal Constructs: Vol. 2. Clinical Diagnosis and Psychotherapy*. Oxford, England: Norton

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A.J., Walters, E.E, & Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association, 289,* 3095-3105.

Kraemer, H. C. , Wilson, G.T. , Fairburn, C. G., & Agras W. S. (2002). Mediators and moderators of treatment effects in randomised clinical trials. *Archives of General Psychiatry, 59*, 877-883.

Kratochwill, T. R., Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: randomization to the rescue. *Psychological Methods, 15*, 12-44.

Kuhn (1970). *The Structure of Scientific Revolutions (2nd edition),* Chicago: University of Chicago Press

Kuyken, W., Warren, F. C., Taylor, R. S. et al….. & Dalgleish, T. (2016). Efficacy of Mindfulness-Based-Cognitive-Therapy in prevention of depressive relapse. *JAMA Psychiatry, 73*, 565-574.

Larkin, M., Boden, Z. V. R., & Newton, E. (2015). On the brink of genuinely collaborative care: Experience-based co-design in Mental Health. *Qualitative Health Research, 25,* 1463-1476.

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2017). Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine, 47*, 1000-1011.

Lee, E. C., Whitehead, A. L., Jacques, R. M., & Juilous, S. A. (2014). The statistical interpretation of pilot trials: should significance thresholds be reconsidered. *BMC Medical Research Methodology, 14*, 41.

Lewin, K. (1943). Psychology and the process of group living. *The Journal of Social Psychology*, 17, 113–131.

Lewin, K. (1951). *Field Theory in Social Science*. Chicago: University of Chicago Press.

Leykin, J., & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating Association from bias. *Clinical Psychology Science and Practice, 16*, 54-65.

Longmore, R.J., & Worrell, M. (2007). Do we need to challenge thoughts in cognitive behavior therapy? Clinical Psychology Review, 27, 173-187.

Lorenzo-Lucas, L. (2018). Representing the heterogeneity of depression in treatment research. *Acta Psychiatrica Scandinavica,* https://doi.org/10.1111/acps.12914

Lorenzo-Luaces, L., German, R. E., & DeRubeis, R. J. 92015). Its complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review, 41*, 3-15.

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders, 213*, 78-85.

May, C. R., Cummings, A., Girling, M., Bracher, M., Mair, F. S., May, C. S., Murray, E., Myall, M., Rapley, T., & Finch, R. (2018). Using normalization process theory in feasibility studies and process evaluations of complex health interventions: a systematic review. *Implementation Science, 13*, 80.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and Structure of Behaviour*, Holt and company, New York.

May C. (2013) Towards a general theory of implementation. *Implementation Science, 8*, 18.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Science, 7*, 141-144.

Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, Moore L, O'Cathain A, Tinati T, Wight D, Baird J. *Process evaluation of complex interventions: Medical Research Council guidance*. MRC Population Health Science Research Network, London, 2014

Mortensen, C. R., & Cialdini, R. B. (2010). Full-cycle social psychology for theory and application. *Social and Personality Psychology Compass, 4*, 53-63.

Morley, S. (2017). Single-case methods in Clinical Psychology: A practical guide. Routledge, UK.

Morris, Z. S., Wooding, S., & Grant, J. (2011)The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*, 510–520.

Munafo, R. M., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J. J., & Ionannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021

Munder, T., Brutsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review, 33*, 501-511.

Nakarada-Kordic, I., Hayes, N., Reay, S. D., Corbet, C., Chan, A., (2017) Co-designing for mental health: creative methods to engage young people experiencing psychosis. *Design for Health, 1*, 229-244.

Nelson, J. C., Delucchi, K. L., & Schneider, L. S. (2013). Moderators of outcome in late-life depression: a patient-level meta-analysis. *American Journal of Psychiatry, 170*, 651-659.

NIHR Trainee Coordinating Centre (2017). Ten years on: Adapting and evolving to new challenges in developing tomorrows health research leaders.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Mellor, D. T. (2018) The preregistration revolution. *Proceedings of the National Academy of Science, 115*, 2600-2606.

Onghena, M., Maes, B., Heyvaert, M.(2018). Mixed Methods Single Case Research: State of the Art and Future Directions. *Journal of Mixed Methods Research*, https://doi.org/10.1177/1558689818789530

Open Science Foundation (2015) Estimating the reproducibility of psychological science. *Science, 349, 6251*, aac4716.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2008). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*, 829-841.

Popper, K. R. ([1974] 2005). *Unended Quest: An Intellectual Autobiography* ([Rev. ed.). London; New York: Routledge

Reardon, K. W., Corker, K. S., & Tackett, J. L. (in press). The emerging relationship between clinical psychology and the credibility movement. *The Behaviour Therapist.*

Rosenthal, R., (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5,* 2-14.

Salkovskis, P. M. (2002) Empirically ground clinical interventions: Cognitive Behavioural therapy progresses through a multi-dimensional approach to clinical science. *Behavioural and Cognitive Psychotherapy, 30,* 3-9.

Saxon, D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: therapy modality, dosage and non-completion. *Administrative and Policy in Mental Health and Mental Health Services Research, 44,* 705-715.

Serghiou, S., & Goodman, S. N. (2018). Random-effects meta-analysis: Summarizing evidence with caveats. *Journal of American Medical Association Guides to Statistics and Methods,* doi:10.1001/jama.2018.19684.

Shafran, R., Clark, D. M., Fairburn, C. G., Arntz, A., Barlow, D. H., Ehlers, A., Freeston, M., Garety, P. A., Hollon, S. D., Ost, L. G., Salkovskis, P. M., Williams, J. M. G., & Wilson, G. T. (2009). Mind the gap: Improving the dissemination of CBT. *Behaviour Research and Therapy, 47,* 902-909.

Shearer, J., & Byford, S. (2015). The basics of economic evaluation in mental healthcare. *BJPsych Advances, 21,* 345-353

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Perspectives on Psychological Science, 22,* 1359–1366.

Simon, H. A., & Newell, A. Human problem solving: The state of the theory in 1970. *American Psychologist, 26,* 145 – 159.

Skinner, B. F. (1955). A case history in scientific method. *Address of the President at the Eastern Psychological Association meetings in Philadelphia,* April 1955.

Skinner, B. F. (1949). Are theories of learning necessary? *Psychological Review, 57,* 193-216

Smaldino, P., Turner, M. A., & Contreras Kallens, P. (2019, January 28). Open science and modified funding lotteries can impede the natural selection of bad science. https://doi.org/10.31219/osf.io/zvkwq

Tackett, J., Brandes, C., King, K., & MArkon, K. (in press). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology.*

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., Shorut, P. E. (2017) It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742-756.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismalia, A. et al (2010). A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology, 10,* 1

Tuffaha, H. W., Gordon, L. G., Scuffham, P. A. (2014). Value of information analysis in healthcare: A review of principles and application. *Journal of Medical Economics, 17,* 377-383.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., Rosenthal, R. (2008). Selective publication of anti-depressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*, 252-260.

Üstün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., & Murray, C. J. (2004). Global burden of depressive disorders in the year 2000. *The British Journal of Psychiatry*, *184*, 386-392.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity and progress. *Perspectives on Psychological Science, 13*, 411-417.

Vittengl, J. R., Clark, L. A., Dunn, T. W., & Jarrett, R. B. (2007). Reducing relapse and recurrence in unipolar depression: A comparative meta-analysis of cognitive-behavioral therapy's effects. *Journal of Consulting and Clinical Psychology*, *75*, 475-488.

Wagenmakers, E. J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science, 13*, 418-427.

Wai, J. & Halpern, D. F. (2018). The impact of changing norms on creativity in psychological science. *Perspectives on Psychological Science, 13*, 466-472.

Wallis, C. J. D., Detsky, A. S., & Fan, E. (2018). Establishing the effectiveness of procedural interventions. The limited role of randomized trial. *Journal of the American Medical Association, 320*, 2421-2422

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin, 134*, 163-206.

Whewell, W. (1840). *The philosophy of the inductive sciences, founded upon their history* (Vol. 2). London, England: John W. Parker. Retrieved from https://archive.org/details/philosophyofindu01whewrich.

Wiles, N. J., Thomas, L., Turner, N., Garfield, K., Kounali, D., Campbell, J. et al.(2016). Long-term effectiveness and cost-effectiveness for treatment-resistant depression in primary care: follow up of the CoBalT randomised controlled trial. *Lancet Psychiatry, 3*, 137-144.

Zimmerman, M., McGlinchey, J. B., Posternak, M. A., Friedman, M., Attiullah, N., & Boerescu, D. (2006). How should remission from depression be defined? The depressed patient's perspective. *American Journal of Psychiatry, 163*, 148-150

Zwann, R. A., Etz, A., Lucas, R. R., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioural and Brain Sciences*, e120.

Dunn et al "A commentary on research rigour in clinical psychological science: How to avoid throwing out the innovation baby with the research credibility bath water in the depression field"

Highlights

- We discuss possible risks if research rigour principles are applied too rigidly
- We use research into depression as an illustration
- Theoretical innovation could be stifled
- A focus on reducing type I error could inadvertently exacerbate type II error
- Treatment development and implementation could be neglected