Draft accepted for publication in *Biology and Philosophy*

From Depressed Mice to Depressed Patients: A Less "Standardized" Approach to Improving Translation

By Monika Piotrowska

We study suffering because alleviating it is good, and inflicting it is wrong. Yet we can't study suffering in an animal that can't suffer, and we can't study suffering without inflicting it.

-Garner 2020, 82

1. Introduction

Depression[1] is a widespread and debilitating disorder, with significant social and economic impacts (World Health Organization 2021). The COVID-19 pandemic has exacerbated the situation, leading to a sharp increase in cases of depression worldwide (Santomauro et al. 2021). Despite the urgent need for effective treatments, developing new antidepressant drugs has proven to be difficult, with roughly 90% of compounds that work on animals failing to work on humans (Garner 2014). To put the point differently, a big problem for developing more effective treatments for depression is the difficulty of extrapolating from seemingly effective treatments used on animals to effective treatments for human beings. One obstacle to this sort of extrapolation is the fact that depression occurs much more frequently among women than men—by an almost 2:1 ratio—and animal research on depression has been done almost exclusively using male animals (Beery and Zucker 2011; Shansky 2019). In light of this, the seemingly easy solution would be to improve the male-female sex ratio in animal research. But such a remedial step does little when the problem itself—i.e., the problem of effective extrapolation—arises from

[1] Although I focus on depression (major depressive disorder), much of my argument generalizes and applies to other mental health disorders.

a deeper theoretical commitment—*viz.,* the commitment to standardization. It is because scientists need standardized animals to control variation that they end up focusing on, for example, only one sex. This theoretical commitment results in female subjects being excluded from animal research (their estrus cycle is a potential source of variability). It is the drive toward this commitment that I aim to engage here. More concretely, my central aim is to argue that seemingly reasonable standardization choices in behavioral neuroscience research on depression often hinder extrapolating meaningful data from rodents[2] to humans. And since poor generalizability of research findings contributes to needless animal waste, standardization raises both scientific and ethical concerns.

The paper opens by discussing the challenges of modeling depression in rodents. In Section 2, I argue that these difficulties help explain why behavioral neuroscience research on depression is highly standardized. In Section 3, I describe a popular *behavioral test* for depression used on rodents and argue that, in light of alternative interpretations, the typical interpretation of this test faces serious challenges. I turn away from the behavioral test in Section 4 to examine ways in which standardizing *test subjects* impedes extrapolation, arguing that since most behavioral tests for depression are done on one inbred strain of rodents, the tests are measuring the norm of reaction for a single, and perhaps, unusual, genotype. Section 5 turns to the *testing environment*, arguing that standardization of housing conditions for rodents can be a source of stress, which complicates extrapolation efforts. Indeed, since humans almost always have some power to influence environmental stressors, testing potential therapies on animals that lack that power makes extrapolation difficult. In the final section, I introduce a thought

[2] Both mice and rats are used in the research I'll be discussing, and both have been standardized by preferring male rodents. Rather than awkwardly discussing 'mice and rats' I'll use 'rodents' unless the distinction matters and despite the fact that 'rodents' includes animals other than mice and rats.

experiment designed to identify unreasonable standardization choices and explain how animal ethics committees can contribute to improving extrapolation from rodents to humans.

2. Modeling Depression in Rodents

Standard practice in biomedical research requires testing the safety and efficacy of drugs on nonhuman animals before enrolling human patients in clinical trials.[3] Various pragmatic factors go into deciding which nonhuman animal ought to be used, but generally an animal makes a *good* model for purposes of research if it shares mechanistic features with the target (cf. Craver & Darden 2013), because the same causal pathways are likely to generate the disease of interest and be affected in similar ways when those pathways are disturbed or repaired through one or another intervention.

When it comes to using rodents as models of human depression, however, the underlying mechanisms are poorly understood. Although researchers know that "serotonin, HPA, neuropeptide, neurotrophin, endocannabinoid, and neuroinflammatory mechanisms" (Gonda et al. 2018, 2) have some influence on depression, much remains unknown. The upshot is that the justification for using rodents to study human depression—*viz.* their shared causal pathways—is largely missing. Consider, for example, that although antidepressants were first introduced almost 70 years ago, currently "more than 50% of patients do not respond to the first treatment they are prescribed and around 30% fail to respond even after several treatment attempts"

[3] This trend is starting to change. For example, in the U.S., the FDA Modernization Act 2.0 was signed into law in 2022, lifting the animal testing requirement. This means that the Food and Drug Administration (FDA) no longer requires all drugs to be tested on animals before human trials. However, given the advantages of testing drugs on living organisms and the fact that the law doesn't *require* the use of alternative methods, skeptics remain unconvinced that the FDA approval process will change significantly in light of the new law (cf. Wadman 2023).

(Menke 2019, 101). Moreover, even newly developed drugs show a high failure rate in clinical trials (Gururajan et al. 2019, 686), and low success rates are often a sign of inadequate understanding of the underlying biological mechanisms of a disease.

The heritable aspect of depression is also poorly understood. For example, no genetic variant has been identified as substantially increasing the risk of depression even though genome wide association studies have identified several common genetic associations of low penetrance[4] (Wray et al. 2018). When modeling diseases in which many genes contribute a small fraction to the genetic risk of developing a disease, the environment, and even the genetic background of the rodent, are more likely to dominate the effects of transgenic manipulation.[5] Thus, even though genetic factors seem to play an important role in the development of depression—as indicated by family, twin, and adoption studies (Sullivan et al. 2000)—genetic research has been of limited use for finding treatments for depression. Indeed, the National Institute of Mental Health (NIMH 2019) removed the 'genes' unit of analysis from its 2007 matrix, reinforcing the point that animals used in research do not have to have genes associated with depression to count as models of depression. As Falk Lohoff (2010) writes, "[d]espite intensive research during the past several decades, the neurobiological basis and pathophysiology of depressive disorders remain unknown" (359).

Slow progress in understanding the biophysiological development of the disease has led researchers to essentially black-box the internal mechanisms of depression and turn instead to

[4] Penetrance of a disease-causing gene is the proportion of individuals carrying it that also exhibit clinical symptoms of the disease. Hence, if a gene has high penetrance, many if not most of its possessors will develop symptoms of the disease. Conversely, if a gene has low penetrance, many individuals will possess the gene but show no symptoms of the disease.

[5] I do not mean to suggest that transgenic mutations cannot yield an effect. There are transgenic mouse lines that show significant depressive-like behaviors, but the translation to humans is often questionable.

focusing on the external causes of the disease—e.g., stress. But this focus introduces its own difficulties. As Tracy Bale and colleagues (2019) explain:

> [P]sychosocial stress represents the major antecedent of depression and other affective disorders. *Therefore, stress is a definable trigger of the presentation of these diseases in vulnerable individuals*…At their core, affective disorders are disorders of stress coping. These disorders arise when the strategies that the organism relies upon to meet environmental challenges are insufficient, dysregulated, or otherwise maladaptive… (1350, emphasis original)

Consequently, studies that aim to model human depression almost always expose rodents to some environmental stressor. Vulnerable individuals would thus seem to be the appropriate target of investigation since they are likely to become depressed in response to an environmental trigger. Even so, it's not clear what makes them 'vulnerable', and they tend to make up only about one-fifth of those exposed to major stressors (Gonda et al. 2018).

Shifting the focus from shared biophysiological mechanisms to shared external inputs (and corresponding outputs) is not unusual in biomedical research. According to Jessica Bolker (2009), animals used as "surrogates" for human patients in clinical trials are often used for diagnostic purposes. The idea is to invoke symptoms in the rodent model, find an effective treatment for those symptoms, then use similar treatments on human targets diagnosed with similar symptoms. Such a method does not require researchers to understand the mechanisms of a disease. As Bolker explains, "where the primary objective is to alleviate symptoms (as may be the case where underlying causes are unknown, intractable or diverse), the model need only

replicate the symptoms" (2009, 490). Thus, it isn't unusual for researchers to use stress as input for causing depression in mice even if they do not fully understand the way in which stress causes depression. Even so, such a diagnostic approach to the problem—i.e., one wherein diagnosing shared symptoms of depression justifies shared treatments between rodents and humans—requires some gauge to determine whether the stressor is giving rise to comparable symptoms in rodent models and human targets.

This raises another difficulty. Determining whether a rodent exposed to one or another stressor is "depressed" relies on an anthropocentric notion of depression, which has a variety of symptoms, some of which have no clear corollaries in rodents or cannot be measured in rodents. According to the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (2022), a diagnosis of depression in humans requires five or more symptoms to be present during a 2-week period. Of these five, the primary symptom of depressed mood and/or loss of interest or pleasure must be on the list. Other secondary symptoms include weight loss or gain, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue, feelings of worthlessness or excessive guilt, decreased concentration, and thoughts of death/suicide. These symptoms must cause clinically significant distress or dysfunction in one's life and must not be explicable by the effects of other medications, substances, or some other medical condition.

Given the diagnostic criteria for humans, researchers working with rodents face a litany of problems when gauging whether rodent models display symptoms of depression with human corollaries. First, researchers simply cannot assess whether a rodent has recurrent thoughts of death or excessive thoughts of guilt. Second, some symptoms may be evidence of not only depression but of several other psychiatric and neurological disorders. For example, according to Susan Stanford (2017), "a deficit in social interaction/social withdrawal is evident in depression,

social phobia, schizophrenia and autism" (642). And, finally, clusters of symptoms—which, after all is what is required for a diagnosis of depression in humans—complicate interpreting symptoms in rodents. For example, two rodents may exhibit different clusters of symptoms and those clusters may run in opposite directions. One rodent might have insomnia, weight loss, and psychomotor agitation, while the other has hypersomnia, weight gain, and psychomotor retardation (Nestler and Hyman 2010). The point is simply that gauging symptoms of depression caused by environmental stressors is not particularly rigorous.

To summarize: modeling depression in rodents using genetic markers has proven difficult. As a result, researchers have black-boxed the potential biological mechanisms, focusing instead on experimental designs that emphasize symptom-producing stressors.[6] Even so, we don't truly understand the relation between stressor inputs and symptoms of depression, which leaves researchers without a reliable metric to evaluate the legitimacy of their experimental designs. To address this concern of reliability, researchers have placed a high value on standardization, which is where I turn now.

Standardization involves controlling aspects of experiments so that, ideally, nothing else can account for a significant outcome except the therapy being investigated.[7] While eliminating all potential confounding variables has been shown to be an unachievable ideal (see Crabbe et al. 1999), eliminating most confounds increases the odds of being able to measure the effect

[6] By focusing on experimental practices that black-box potential mechanisms, my work diverges from most philosophical accounts of neuroscience (cf. Bechtel 2008; Bickle 2003; Craver 2007). Works by Jacqueline Sullivan (2009) and Nina Atanasova (2015) are notable exceptions.

[7] Standardization is also prevalent in human clinical trials. For instance, many clinical trials exclude patients with potentially confounding variables such as elderly patients, patients with co-morbidities, and those taking additional medications, thereby excluding patients who are most likely to use the medication once it becomes available (cf. Bluhm and Borgerson 2018). However, while such exclusions undermine the generalizability of research findings, the effects of standardization are not as pronounced as in preclinical trials since human clinical trials allow for considerably more variation than animal experiments.

produced by the intervention of interest. Hence, in the early stages of investigation, it is standard practice in animal research to eliminate potentially confounding variables by using highly similar test subjects maintained in highly similar environments. But in the behavioral neuroscience of depression, strict standardization practices are common even in later, preclinical stages of research in part because depression is poorly understood, which motivates the fear that significant findings will not replicate unless everything is kept the same (cf. Garner 2014). Although philosophers have been rightly critical of treating replication as an overarching epistemic value in science (cf. Guttinger 2020; Leonelli 2018), the fact remains that in the behavioral neuroscience of depression, "the ability to obtain the same results through the repeated application of the same research methods" (Leonelli 2018, 135)—what Leonelli calls 'direct reproducibility'—continues to be the gold standard. If an attempt to replicate a significant result fails, it suggests that the original finding was spurious, that the researcher did not properly understand the experiment's main causal factors, the ways they interact, or the phenomena they bring about. Keeping everything the same through standardization serves to ensure that the model will not "break" when the experiment is repeated, even if researchers do not fully understand the results.

However, standardization also narrows the range of targets analogous to the model. When an experimental result has only been true in a very particular, highly standardized context, the ability to extrapolate information about cases outside the standardized context diminishes (cf. Sullivan 2009; Würbel 2000). As a result, the experiment risks turning into a single case study, with limited information to be inferred about entities external to the experimental model, e.g., human patients. In the next three sections, I delve into the costs of standardization for extrapolation in greater detail by examining three aspects of behavioral neuroscience

experimental designs that investigate depression: behavioral tests, test subjects, and testing environments. In each section, I demonstrate how seemingly reasonable choices about standardization can serve as obstacles to the possibility of extrapolating findings to the human population.

3. The Behavioral Test

A popular behavioral test used to screen treatments for depression in rodents is the Porsolt Forced Swim Test (FST) (Porsolt et al. 1977). The FST uses an aversive stimulus to induce an animal's fear of death, motivating behaviors aimed at escaping the stimulus, and, ultimately, leading the animal to despair. As Jacqueline Crawly (2007) notes, "[f]ear of death…drives a great deal of behavior," but "[w]hen the aversive stimulus is inescapable, the animal will eventually stop trying to escape" (231). Behavioral tests that use inescapable aversive stimuli to motivate behavior—such as the FST—are called *behavioral despair* tests. Although all animals will eventually give up the fight—i.e., all animals will eventually succumb to despair under the right aversive conditions—behavioral despair tests operate on the assumption that depressed rodents will succumb earlier than nondepressed rodents. That is, a depressed animal will give up fighting for survival earlier than one that is not depressed. On the basis of this assumption, researchers use behavioral despair tests to screen animals for depression. Those that succumb to despair sooner are sorted as depressed and those that continue fighting are set aside as not depressed.

The FST is the most widely used behavioral despair test for rodents. It involves placing a rodent in a tall clear cylinder filled with room temperature tap water. The water is deep enough

to prevent the rodent from balancing on its feet or tail, and the cylinder extends far enough above the upper surface of the water to prevent the animal from climbing out of the cylinder. Rodents will generally swim to find an escape route, but eventually they'll stop swimming and float on the surface of the water. The test measures the time spent swimming versus the time spent floating, typically during a 10-minute session. The standard interpretation is that "[f]loating time is considered the measure of depression-like behaviors, in that the animal has stopped swimming and 'given up' on finding an escape route" (Crawley 2000, 140). Hence, rodents that float sooner are considered more depressed.

The appeal of the FST is that it matches our expectations of depression. We expect a depressed individual to give up sooner because they are likely to lack motivation to continue in the face of adversity. But even if this assumption is correct, there has been no research to show its legitimacy for humans in conditions similar to those of rodents—i.e., it has never been shown that depressed humans stop swimming and succumb to despair sooner than their nondepressed counterparts. Further, alternative explanations for the observed rodent behavior are obviously available: perhaps the rodents who gave up sooner are simply more intelligent, realizing the futility of their efforts more quickly than their persistent counterparts. Whatever the explanation, we cannot unproblematically infer that rodents who swim less and float more are depressed.

Stanford (2017) makes a similar observation regarding a related behavioral despair test, called the Tail Suspension Test (TST), in which a rodent is hung upside down by its tail:

> It is often claimed that a deficit in struggling in the Tail Suspension Test reveals depression-like behavior, but if two humans were suspended upside-down, against their

> will, I doubt a difference in their struggling would be attributed to one being more depressed than the other. (641)

But why would we resist attributing depression to humans in this imagined scenario? Plausibly, because there are indefinitely many alternative explanations that would be similarly warranted. If that is so, we have reason to be skeptical of any particular explanation as the right one. This criticism extends to interpretations of the FST as a legitimate test to screen for depression (See Box 1 for alternative interpretations of what the FST is measuring). As Stanford reminds us, there are at least twenty diagnostic symptoms of depression in humans, but "a motor deficit…when confined within a tank of water" (2017, 641) is not one of them.

A slightly less controversial use of the FST screens for depression after treating rodents with antidepressant drugs. Here, rodents are subjected to the FST, antidepressant drugs are doled out, then rodents are subjected to the FST again. Comparing the swim time between events is then used to determine the efficacy of the drug for preventing despair. Such treatments have been shown to reduce total float-time in rodents and bolster their escape behavior. Interestingly, these positive outcomes in rodents have been successful at predicting human responses to antidepressant drug treatments (Castagné 2011)—i.e., drugs that reduce float time and bolster escape behavior predictably produce positive anti-depressant effects in humans.[8] Of course, one might worry that just because antidepressants increased the duration of escape behavior does not mean that the rodents were depressed prior to treatment, which is to say that the drugs didn't obviously have an antidepressant effect. In fact, explaining the effect as "antidepressant" might be the result of what Jennifer Radden calls *drug cartography*, "a remapping of psychiatric

[8] Although this long-held assumption has been scrutinized in recent years (cf. Trunnell and Carvalho 2021).

categories based not on traditional symptom clusters but on psychopharmacological effects" (2003, 38). The drug cartography trend made it acceptable for practitioners to classify depression as any condition alleviated by antidepressants.

Radden (2003) points to several problems with classifying depression in this way. Even so, behavioral neuroscientists do not have to rely on drug cartography to justify their use of the FST. Rather, if the FST is an effective predictor of the efficacy of antidepressants for humans, rodents that swim longer post-treatment plausibly bear some similarity to depressed humans for whom antidepressants work effectively in alleviating symptoms of depression.[9] The persuasiveness of this inference has led the FST to become the standard behavioral test for investigating the efficacy of new drug therapies for depression. Although it was invented in the 1970s, its popularity grew in the 2000s when researchers started genetically modifying rodents to induce depression. The idea was that the FST offered a standardized (as well as quick and easy) way to determine whether a given genetic modification actually induced depression. By 2015, researchers were publishing an average of one paper per day using the FST (Reardon 2019, 456-7) as a reliable measure of rodent depression, and its popularity does not appear to be declining (cf. Petkovic and Chaudhury 2022).[10]

Given its ease of use and reliability across laboratories, it's not surprising that the FST became the standardized method for screening for depression in rodents. What is surprising, however, is that its popularity has not waned despite several alternative interpretations that raise serious doubts about what the FST is actually measuring (see Box 1). The upshot is that we seem

[9] One problem with this inference is that there is a mismatch between the amount of time it takes for the antidepressants to work in humans in contrast to rodents. When testing rodents, responses in the FST are elicited after a single dose, whereas "the human antidepressant response to these medications occurs only with chronic (several weeks to months) administration, suggesting that the mechanism(s) by which they alter behavior in the…FST are distinct from their antidepressant action in humans" (Bale 2019, 1351).

[10] Although the popularity of the FST has not waned among researchers, it has waned among pharmaceutical companies following a recent campaign by PETA (cf. PETA 2022).

to have standardized a method of screening for depression even though that method doesn't unproblematically screen for depression. Given this fact, we might be impeding our efforts to extrapolate from rodents to humans because we have standardized the wrong behavioral test. As Andres Hånell and Niklas Marklund (2014) remind us, "In the interpretation of behavioral test results, an important issue is to understand the cause of the observed behavior" (5). For over fifty years, the standard interpretation of the FST has been that depression is the cause of immobility, but several alternative interpretations raise doubts about that interpretation.

4. The Test Subject

Let us now turn to the standardization of rodent models and the impediment it poses to extrapolating significant results for human beings. When researchers run behavioral tests such as the FST, the test subjects themselves—i.e., the rodent models—have been genetically standardized. The idea is that when rodents are genetically homogenous, differences in experimental outcomes won't be explainable in terms of individual genetic differences. Instead, outcomes will be explained by something else—ideally, by the therapies being tested. For mice, the standardized mouse-type used to populate the homogenous test groups is typically Black-6 (formally known as C57BL/6). Its genome was sequenced early and, as a result, many knockout and transgenic experiments have been done against its genetic background. Since the strain has been used for some time, it has a far greater number of generations of brother and sister (or parent and offspring) matings than other strains of mice meeting the requirements for qualifying as inbred. Consequently, Black-6 mice are essentially homozygous at all loci (Beck et al. 2000).

Because of their ubiquitous use in research, Black-6 mice have come to represent both the typical mouse and the typical human—the latter by virtue of being the most widespread model used to extrapolate information about them. This is problematic for at least two reasons. First, compared to other strains of inbred mice, Black-6 tends to be an outlier. For example, Jeffrey Mogil et al. (1999) tested 11 different inbred mouse strains and ran them through a dozen standard tests for pain sensitivity. No inbred strain was more unusual in its response to painful stimuli than Black-6. Further peculiarities about Black-6 include that it is one of the only strains of mice that drinks alcohol voluntarily,[11] it is prone to morphine addiction, and has several other quirks. Daniel Engber (2011) describes Black-6 as "a teenaged, alcoholic couch potato with a weakened immune system, [that] might be a little hard of hearing." Hence, there is a legitimate worry that we are running behavioral tests on a strain of mouse that is not particularly representative of mice, let alone human beings.

Second, even if researchers chose a different strain of rodent with different traits, the strain would still fail to represent the heterogenous human population. Recall that in the human population, only a small fraction of vulnerable individuals become depressed following a stressful event. The inference from this fact is that only some genotypes are susceptible to stressor-induced depression. But such genotypic differences are ignored when researching depression in rodents. When rodents are stressed in the lab, all of them—as a genetically homogenous group—swim less and float more in the FST than rodents in the control group that have not been stressed. Hence, there is a significant disanalogy between the model rodent population and the target human population. If the rodent population were representative of the human population, only a small fraction of the stressed rodents would swim less and float more.

[11] Most strains of mice will not drink alcohol voluntarily. For a fascinating account of the difficulties involved with getting mice to drink alcohol in order to study alcoholism in humans, see Chapter 5 in Nelson (2018).

Thus, by using a single inbred strain of rodents, such as Black-6, researchers are not modeling the complex gene-environment (GxE) interaction that results in only some people becoming depressed. Instead, they are measuring the norm of reaction for a single, and perhaps unusual, genotype.

Not only are most rodent test subjects of a single genetic strain, until recently, they were also of a single, male sex,[12] which in humans is the sex half as likely to become depressed as the other. The inclination to use only male rodents in experimental studies raises obvious difficulties for extrapolating to human populations that are not uniformly male. Even so, some researchers (e.g., Johnson et al. 2021; Shansky and Murphy 2021) have cautioned against the new inclusionary requirements because simply running females through existing behavioral tests fails to consider sex-biases built into the experimental apparatus.

Consider two ways standardized behavioral tests might have a built-in sex bias. First, when testing for multi-symptom diseases like depression, where some symptoms are more common in a particular sex, tests that rely on those symptoms for determining expression of the disease may pick it out only in the one sex. According to Alyssa Johnson and colleagues (2021), some symptoms of depression are sex specific: "[w]omen report a higher occurrence of appetite-related issues, loss of interest, and suicidal thoughts…while men report higher rates of aggression and suicidal thoughts" (73). If this difference in symptoms typical of human beings is also present in rodents, researchers relying on sex-specific symptoms to diagnose rodent depression might fail to observe depression in the portion of the population that doesn't express the disease in that manner. Presumably, then, researchers ought to make sure their behavioral

[12] This changed in 2016 when the U.S. National Institutes of Health introduced a mandate requiring grant recipients to use both sexes in animal studies (Clayton 2016). Similar policies had already been implemented by the Canadian Institutes of Health Research and the European Commission.

tests can pick out multiple symptoms or that they are using more than one behavioral test to screen for depression.

A second way in which standardized behavioral tests might discriminate by sex is by relying on assumptions about the "right way" to solve a problem. Consider, for example, the Morris Water Maze Test used to measure spatial learning in rodents. When researchers place rodents in a water maze, their spatial navigation is assessed by measuring the speed at which the test is performed. Since males tend to learn to navigate quicker than females, the standard interpretation is that females are spatially impaired in comparison to males. But Rebecca Shansky and Anne Murphy (2021) argue that what appear to be quantitative sex differences may reflect qualitatively different, sex-dependent strategies. Indeed, this turned out to be true for the water maze test. Females were traveling farther and thus taking longer to complete the test because they were approaching the task differently. Here is how Shansky and Murphy (2021) explain this outcome:

> [F]emales were using a qualitatively different strategy, one that, although more circuitous, minimized exposure to predators and other dangers. Subsequent studies further showed that sex differences in time to complete a water maze task were completely eliminated if the animals had prior exposure to the maze. This example illustrates the need to reevaluate the biases inherent in our experimental designs regarding the 'right' way to solve problems and ask whether what appear to be errors in fact simply reflect the selection of a different strategy. (458)

If there are qualitative sex differences in how rodents approach behavioral tests, assessments based on the same metrics—e.g., speed—risk misinterpreting the outcome as a difference in ability. The upshot is that adding female test subjects to standardized experimental designs may require doing more than simply assessing how they perform on behavioral tests designed for males.

In sum, the way in which test subjects have been standardized in the behavioral neuroscience of depression has implications for what can and can't be extrapolated from the research. Extrapolating to a heterogenous human population is difficult when behavioral tests measure the norm of reaction for a single, and perhaps unusual test subject like Black-6. Similarly, standardized tests may not apply to test subjects that have recently been mandated for use, like female rodents, because of existing sex-biases built into experimental design.

5. The Testing Environment

As a segue to looking at the ways in which standardized testing environments might also undermine extrapolation efforts, consider one argument that has been made for ending the exclusion of female rodents from experimental studies. A meta-analysis revealed that male test subjects displayed cyclical behavioral variations similar to female counterparts. Thus, excluding female rodents was based on the mistaken assumption that their estrus cycle would introduce variability not present in male rodents. As Anand Gururajan et al. (2019) reported:

> [M]eta-analyses of preclinical data have shown that the variance in data obtained from cycling female rodents does not differ from that obtained in males with regard to a range

> of behavioral, physiological or molecular traits. Indeed, the factor that most strongly influenced data variance was housing conditions. (694)

In fact, analyzing existing research reveals that several environmental factors add more variability to behavioral outcomes than sex-differences, including the position of the animal's cage (whether it was at the top or the bottom of the shelving units (Izidio et al. 2005)), the sex of the person who conducted the experiment (Sorge et al. 2014), and the lighting conditions of the housing unit (Richetto et al. 2019).

For most behavioral neuroscientists, these findings are not news. Researchers have been well-aware of confounding environmental conditions adding unwanted variability to test results (see, for example, Crabbe et al. 1999). Indeed, after a year of observation in a behavioral neuroscience lab, Nicole Nelson (2018) wrote that:

> Researchers treated behavior as something . . . that could change depending on the details of the experimental protocol, the time of day, or the person conducting the experiment. It felt as though researchers were living in an extended moment of doubt and uncertainty (11).

The uncertainty imposed on experimental results by features of the testing environment is made worse when we realize that rodents are affected by variables which we do not even experience. For example, "animals see color we do not, hear sounds we do not, have electrical and magnetic senses that we do not, respond to odors and pheromones that we can't detect" (Garner et al. 2017, 109). We simply can't understand how such variables affect animal psychology. And

further, even factors that influence results of which we are aware, and which may have a detectable influence on test results, are not always reported in the literature—e.g., season and humidity.

But even if we had better reporting, it likely wouldn't resolve the difficulties imposed by environmental confounders. Indeed, Hanno Würbel (2002) has argued that better reporting is not the solution. He writes, "by pretending 'to list all factors that affect mouse behavior,' such lists may in fact divert attention away from highly relevant factors that were not considered, were considered to be irrelevant, too difficult to assess, or simply cannot be listed" (5). For example, in a series of studies in the late 1990s (Würbel et al. 1996; Würbel at al. 1998), Würbel and colleagues showed that rodents housed under standard laboratory conditions were more likely to develop *stereotypies*, which are abnormal, repetitive behaviors such as repeated back flips or gnawing on cage bars. This behavior was determined to be the result of the cages holding the animals. The problem, it was determined, was that standard cages lacked enrichment, which resulted in animals behaving abnormally. Here is how Joseph Garner (2014) captures the importance of enrichment:

> Animals are fundamentally designed to control stressors that they care about. Indeed, animals that cannot control (through behavior or physiology) even innocuous stressors can show catastrophic changes in biology. Thus we can define stress as the state in which an animal can safely control a stressor, and distress as the state in which an animal can control a stressor only by negatively impacting another biological system. For instance, if mice find standard housing conditions aversively cold, they can use nesting material to control this stressor; and without nesting material, they are demonstrably distressed as

> their reproductive output suffers. Thus the fundamental argument for enrichment is that an animal is actually abnormal without it. (450)

Testing potential therapies on distressed animals is unlikely to translate well to humans whose living conditions are enriched. Indeed, if standard cages cause this sort of abnormal behavior, what could researchers hope to extrapolate about humans from research involving rodents housed under such conditions? As Barry Yeoman (2003) observed, scientific experiments may not be worth much if they are done on animals that "may be out of their minds" (64).

But making decisions about how best to enrich the testing environment to minimize stress can be quite complicated. Consider the question of whether rodents should be housed together or separately (Kappel et al. 2017). On the one hand, rodents are social animals, so housing them in groups is good for their welfare. It is also good for researchers because housing them in groups allows for modeling the effects of social support on human disease. When discussing the finding that rodents housed in groups are more responsive to chemotherapy, Garner (2014) writes, "these effects are not confounds—they actually mimic the profound impact that social support has on disease progression in humans" (451). But researchers also know that male rodents are not good at sharing territories—which they are forced to do when confined to a cage—and that the stress of repeated social defeat and subordination can diminish their welfare. Hence, deciding how the testing environment is to be standardized (should rodents be housed individually, in pairs, or in groups?) can be a difficult task, producing unknown effects on experimental results.

Another difficulty that arises when considering environmental confounders is that researchers may not even know that their rodents are stressed. For example, the ideal temperature for rodents—the environmental temperature at which rodents do not have to generate or lose

heat—is close to 30 degrees Celsius. However, lab rodents are typically housed at temperatures between 19 and 22 degrees Celsius, which happens to be a temperature range comfortable for clothed humans. This sub-optimal housing temperature means that rodents must use metabolic activity for heat generation, which in turn reduces the energy available for other metabolic and biological functions (Vialard and Olivier 2020). Hence, the task of improving model reliability—i.e., improving the potential to extrapolate meaningful results from animal models—requires an awareness of standard lab conditions, how they might be stressing the animals, and how those stressors affect experimental results.

Of course, humans do not live in stress-free environments. Even so, we typically have some control over stressors. By removing an animal's ability to control their stressors by insisting on standardization of the animal's housing environment, researchers might be testing therapies on animals in a setting that is disanalogous to the one in which humans find themselves. The upshot is that these seemingly reasonable choices about which environmental conditions to standardize might actually be hindering our ability to extrapolate findings to the human population.

6. A Way Forward

Depression is a prevalent disease, but progress in discovering effective treatments has been underwhelming, with only one in nine drugs that enter human trials achieving success. There are various plausible reasons for this failure—e.g., small sample sizes, subjective bias, and improper statistical analysis—but I have focused on explanations involving standardization

practices in animal research, which have received considerably less attention. Put simply: my worry is that highly standardized conditions undermine extrapolation efforts.

The issue at hand is part of the persistent tradeoff between strategies to ensure internal and external validity of animal models in translational biomedical research. Sullivan (2009) characterizes this conflict as a tradeoff between simplicity and complexity. Internal validity focuses on simplifying experiments to guarantee reproducibility of laboratory effects, while external validity emphasizes introducing complexity into experiments to ensure greater resemblance between laboratory models and their human targets.[13] However, in the context of behavioral neuroscience research on depression, internal validity has been prioritized at the expense of external validity. This mistaken prioritization is referred to as the standardization fallacy by Würbel (2000), which is the mistake of increasing reproducibility "at the expense of external validity" (263). I too have argued that highly reproducible results performed under highly standardized conditions are unlikely to generalize. By standardizing (1) the tests used to diagnose symptoms of depression in rodents, (2) the rodents themselves—i.e., the subjects of the experimental tests—and (3) the testing environment, researchers have been learning about depression using experiments that are eerily alike and eerily unlike what they aim to model (Garner et al. 2017).

Imagine implementing the standardization practices used in rodent models on human beings. Suppose we set out to study depression and the effects of antidepressants using only genetically homogenous males who were living in studio apartments with identical furniture, thermostats set to uncomfortably cold temperatures, and working identically monotonous jobs while eating the same boring food. Imagine further that to determine whether the antidepressant

[13] Sullivan (2009) uses the term ‘reliability’ instead of internal validity, and ‘validity’ instead of external validity.

treatment of our imagined bachelor was working, we placed him (and those of his homogenous population) into a small pool of deep water with walls too high to escape and measured how long he would swim before giving up and turning to float on his back. What would such an experiment tell us about depression in the broader population? Would we really concede that such bachelors were a good model from which meaningful data could be extracted? I don't know, but inferences drawn from such an experiment would be far from unproblematic.

This thought experiment is slightly unfair to animal researchers because it's not obvious what the alternative strategies for investigating depression could or should be. Further, given that we can't perform such experiments on humans, it would be a mistake to require complete parity between rodent models and their human targets. But the point of the imaginative exercise isn't to minimalize animal research, it's to help us see things we have failed to notice. As Garner and colleagues (2017) explain, thinking of animals as patients instead of experimental tools, "forces us to think about all the aspects of the experimental background that differ from humans that we might otherwise ignore" (108). When we imagine running an experiment on a genetically homogenous population of men who live in identically monotonous environments through the problematic forced swim test, the seemingly reasonable standardization practices appear problematic, so much so that they seem to seriously undermine efforts to generalize findings to people suffering from depression.

But my complaint isn't that behavioral tests, rodent models, and testing environments have been standardized in the wrong ways (although in the examples I've chosen, they probably have been). Rather, my complaint is that standardizing animal experiments increases the risk of findings that are only true under narrowly defined conditions, ones that won't generalize. Standardization eliminates extraneous noise, which is a good thing, but it also eliminates other

things, like biological reality. Variation is ubiquitous in biology and attempts to eliminate it rest on the mistaken assumption that there is a single, pure treatment effect that can be measured once all differences have been removed. The fact is that no such effect exists, except for the one generated by the interplay between a particular set of genes and environmental conditions (cf. Voelkl et al. 2020; Lewontin 1984). By removing variation, strict standardization practices are generating limited and localized "truths" (von Kortzfleisch et al. 2022, 3) that provide insights into expected outcomes under narrowly defined conditions.

Even so, the solution is not to get rid of standardization altogether. Doing so would simply reverse the problem, leading us to prioritize external validity at the expense of internal validity. Instead, variation can be added in small increments over time, allowing researchers to detect differences in treatment effects produced by different gene-environment interactions. For example, a researcher may split an experiment into several "mini-experiments" with slightly varying conditions, thereby increasing the inference space of experimental findings (cf. von Kortzfleisch et al. 2020; Voelkl et al. 2020).[14] And if an antidepressant shows promise on a variety of behavioral tests, strains of mice, and environmental conditions, we can feel more confident that the treatment effect does not hinge on some particular set of controls and is more likely to generalize. With every difference that's introduced, the confirmatory power of a replication increases (Schmidt 2016).

Of course, most of the burden of improving extrapolation through added variation—or by other means for that matter—falls on the scientific community. To their credit, they have taken significant steps to improve extrapolation—e.g., by creating guidelines for the planning of animal experiments and improving transparency (cf. National Centre for the Replacement,

[14] The concept of "mini-experiments" shares similarities with Sullivan's (2007) notion of "incremental experiments," as they both aim to introduce small, incremental variations.

Refinement, and Reduction of Animals in Research 2010; Animals in Science Regulation Unit 2014). Be that as it may, broad scale changes to improve extrapolation from preclinical animal trials to human patients likely require restructuring the incentives of those who approve, regulate, fund, and publish scientific research, and ethical considerations provide one such incentive. Indeed, excessive homogeneity in preclinical trials is an ethical issue because it contributes to the waste of animal lives with no offsetting benefit for human beings.

Animal ethics committees enforce the responsible use of animals in research using the familiar guidance of the 3Rs: Reduce, Refine, Replace (Russell & Burch 1959). The Refinement principle dictates that animals should not be harmed unnecessarily. Wasting animals in preclinical research that is unlikely to generalize to human patients violates this principle. Thus, it's within an ethics committee's purview to address the question of the generalizability of findings in the review process.[15] Animal ethics committees can get involved in the collective project of improving extrapolation by drawing attention to the need for generalizability as a goal and setting reasonable expectations as conditions for study approval, and they can do this while deferring to scientists for experimental design. For example, animal ethics committees could ask researchers to discuss their choices regarding the variation they introduced, or didn't introduce, and how it applies to the intended inference space. As Korrina Duffy and colleagues (2020) explain, asking researchers to justify certain choices as part of ethical review does not increase the regulatory burden of review committees but forces researchers to consider the generalizability of their findings early in the research process. Thus, for example, even if researchers are allowed to provide a compelling justification for the lack of variability in their sample, providing that justification forces them to confront the effects of a homogenized study

[15] For an in-depth defense of the claim that it's within an ethics committee's purview to address the generalizability of findings, see Piotrowska (2023).

sample on the generalizability of their findings (and hopefully nudges them to adjust their study design in a way that will improve extrapolation).

Even if the regulatory burden imposed by requesting researchers to discuss variation in their research proposals were less than minimal, animal ethics committees should still accept the challenge. If we continue to downplay the significance of variation for predicting the safety and effectiveness of drugs in human patients, we risk the continued waste of animal lives and unnecessary animal distress for the sake of flawed research. We also risk missed opportunities for better treatment options for people suffering from depression. Animal ethics committees can play a crucial role in highlighting the importance of variation in preclinical experiments and establishing reasonable expectations as prerequisites for study approval. Doing so can contribute to the collective effort of improving translation from suffering rodents to suffering patients.

Acknowledgements:

I am grateful to Damian Zuloaga for inspiring me to write this paper and for sharing his scientific expertise. Thanks to Gunnar Babcock, Steve Downes, P.D. Magnus, Matt Mosdell, Anya Plutynski, anonymous reviewers, and the editor of this journal for valuable feedback. Audience input at the Society for the Study of Ethics and Animals colloquium and the British Society for the Philosophy of Science 2022 conference also improved the paper.

Box 1

Interpreting the Results of a Forced Swim Test

Standard Interpretation:

Immobility, or floating, represents a behavioral manifestation of "despair," while escape-directed behavior, such as swimming, is an indication of the absence of despair. Therefore, rodents that stop swimming earlier are generally considered to be in despair since they have given up on finding an escape route prematurely.

Alternative Interpretations:

1. **Coping:** Immobility represents a passive coping strategy adopted by rodents when they realize escape from the cylinder is impossible (de Kloet and Molendijk 2016). On this reading, the rodents transition from active to passive behavior in order to conserve energy and wait for a possible escape. Thus, it can be argued that rodents that start floating earlier are not necessarily in despair but are instead employing an energy-saving strategy to increase their odds of survival.

2. **Memory:** Immobility in rodents is dependent on learning and memory, and the use of antidepressants may interfere with memory consolidation. There is evidence suggesting that antidepressants can affect the ability of rodents to consolidate the learned immobility response (Molendijk and de Kloet 2015). When rodents are exposed to the forced swim test twice—with the first exposure inducing a stable level of immobility and the second exposure measuring immobility after antidepressant treatment—they may forget that floating earlier is an effective coping strategy for conserving energy.

Therefore, on this interpretation, increased escape-directed behavior does not necessarily indicate the absence of despair but rather may be due to forgetfulness.

3. **Anxiety:** Reduced immobility and increased escape-directed behaviors are driven by anxiety (Anyan and Amir 2018). On this view, rodents that swim longer following exposure to antidepressants are displaying anxiety, whereas rodents that start floating earlier are less anxious. Thus, the length of swimming in the FST may not necessarily reflect the absence of despair, but rather the level of anxiety in the animal.

References

Animals in Science Regulation Unit. (2014). PREPARE guidelines for planning animal research and testing. National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs). https://www.nc3rs.org.uk/sites/default/files/documents/Guidelines/NC3Rs%20PREPARE%20Guidelines%20%28second%20edition%29.pdf

Anyan, J., Amir, S. (2018). Too Depressed to Swim or Too Afraid to Stop? A Reinterpretation of the Forced Swim Test as a Measure of Anxiety-Like Behavior. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, *43*(5), 931–933.

Atanasova, N. A. (2015). Validating animal models. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *30*(2), 163-181.

Bale, T.L., Abel, T., Akil, H. *et al.* (2019). The critical importance of basic animal research for neuropsychiatric disorders. *Neuropsychopharmacol.* 44, 1349–1353.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Taylor & Francis.

Beck, J., Lloyd, S., Hafezparast, M. *et al.* (2000). Genealogies of mouse inbred strains. *Nature Genetics* 24, 23–25.

Beery, A. K., & Zucker, I. (2011). Sex bias in neuroscience and biomedical research. *Neuroscience and biobehavioral reviews*, *35*(3), 565–572.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account* (Vol. 2). Springer Science & Business Media.

Bluhm, R., & Borgerson, K. (2018). An Epistemic Argument for Research-Practice Integration in Medicine. *The Journal of medicine and philosophy*, *43*(4), 469–484.

Bolker, J. A. (2009). Exemplary and surrogate models: two modes of representation in biology. *Perspectives in Biology and Medicine*, *52*(4), 485-499.

Castagné, V., Moser, P., Roux, S., & Porsolt, R. D. (2011). Rodent models of depression: forced swim and tail suspension behavioral despair tests in rats and mice. *Current protocols in neuroscience*, *Chapter 8*. https://doi.org/10.1002/0471142301.ns0810as55

Clayton J. A. (2016). Studying both sexes: a guiding principle for biomedicine. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, *30*(2), 519–524.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science (New York, N.Y.)*, *284*(5420), 1670–1672.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Craver C. F., Darden L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*, University of Chicago Press.

Crawley, J. N. (2000). Behavioral Phenotyping of Transgenic and Knockout Mice: Experimental Design and Evaluation of General Health, Sensory Functions, Motor Abilities, and Specific Behavioral Tests, *ILAR Journal*, Volume 41, Issue 3, 2000, Pages 136–143.

Crawley, J.N. (2007). *What's Wrong With My Mouse?: Behavioral Phenotyping of Transgenic and Knockout Mice*. Second Edition. Hoboken (New Jersey): John Wiley & Sons.

de Kloet, E. R., & Molendijk, M. L. (2016). Coping with the Forced Swim Stressor: Towards Understanding an Adaptive Mechanism. *Neural plasticity*, *2016*, 6503162. https://doi.org/10.1155/2016/6503162

Diagnostic and Statistical Manual of Mental Disorders (2022), Fifth Edition, Text Revision (DSM-5-TR™)

Duffy, K. A., Ziolek, T. A., & Epperson, C. N. (2020). Filling the Regulatory Gap: Potential Role of Institutional Review Boards in Promoting Consideration of Sex as a Biological Variable. *Journal of women's health (2002)*, *29*(6), 868–875. https://doi.org/10.1089/jwh.2019.8084

Engber, D. (2011). The trouble with Black-6. *Slate*. Retrieved 16 December 2021. Url: http://www.slate.com/articles/health_and_science/the_mouse_trap/2011/11/black_6_lab_mice_and_the_history_of_biomedical_research.html

Garner J. P. (2014). The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it?. *ILAR journal*, *55*(3), 438–456.

Garner, J. P., Gaskill, B. N., Weber, E. M., Ahloy-Dallaire, J., & Pritchett-Corning, K. R. (2017). Introducing Therioepistemology: the study of how knowledge is gained from animal research. *Lab animal*, *46*(4), 103–113.

Garner. J.P. (2020). The Mouse in the Room. In: Principles of Animal Research Ethics. Edited by: Tom L. Beauchamps and David Degrazia, Oxford University Press

Gonda, X., Hullam, G., Antal, P. *et al.* (2018). Significance of risk polymorphisms for depression depends on stress exposure. *Scientific Reports* 8, 3946. https://doi.org/10.1038/s41598-018-22221-z

Gururajan, A., Reif, A., Cryan, J. F., & Slattery, D. A. (2019). The future of rodent models in depression research. *Nature Reviews Neuroscience*, *20*(11), 686-701.

Guttinger S. (2020). The limits of replicability. *European Journal for Philosophy of Science*, 10(10).

Hånell, A., & Marklund, N. (2014). Structured evaluation of rodent behavioral tests used in drug discovery research. *Frontiers in behavioral neuroscience*, *8*, 252.

Izídio, G. S., Lopes, D. M., Spricigo, L., Jr, & Ramos, A. (2005). Common variations in the pretest environment influence genotypic comparisons in models of anxiety. *Genes, brain, and behavior*, *4*(7), 412–419.

Johnson, A., Rainville, J. R., Rivero-Ballon, G. N., Dhimitri, K., & Hodes, G. E. (2021). Testing the Limits of Sex Differences Using Variable Stress. *Neuroscience*, *454*, 72–84.

Kappel, S., Hawkins, P., & Mendl, M. T. (2017). To Group or Not to Group? Good Practice for Housing Male Laboratory Mice. *Animals: an open access journal from MDPI*, *7*(12), 88.

Leonelli, S. (2018), "Rethinking Reproducibility as a Criterion for Research Quality", *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise* (*Research in the History of Economic Thought and Methodology, Vol. 36B*), Emerald Publishing Limited, Bingley, pp. 129-146.

Lewontin, R. C. (1984). *Adaptation*. Scientific American Library.

Lohoff F. W. (2010). Overview of the genetics of major depressive disorder. *Current psychiatry reports*, *12*(6), 539–546.

Menke A. (2019). Is the HPA Axis as Target for Depression Outdated, or Is There a New Hope?. *Frontiers in psychiatry*, *10*, 101.

Mogil, J. S., Wilson, S. G., Bon, K., Lee, S. E., Chung, K., Raber, P., Pieper, J. O., Hain, H. S., Belknap, J. K., Hubert, L., Elmer, G. I., Chung, J. M., & Devor, M. (1999). Heritability of nociception I: responses of 11 inbred mouse strains on 12 measures of nociception. *Pain*, *80*(1-2), 67–82.

Molendijk, M. L., & de Kloet, E. R. (2015). Immobility in the forced swim test is adaptive and does not reflect depression. *Psychoneuroendocrinology*, *62*, 389–391.

National Centre for the Replacement, Refinement, and Reduction of Animals in Research. (2010). ARRIVE guidelines. Retrieved from https://www.nc3rs.org.uk/arrive-guidelines

Nelson, N. C. (2018). *Model behavior: Animal experiments, complexity, and the genetics of psychiatric disorders.* University of Chicago Press.

Nestler, E., Hyman, S. (2010). Animal models of neuropsychiatric disorders. *Nature Neuroscience* 13, 1161–1169.

NIMH (2019) NOT-MH-19-053: notice of NIMH's considerations regarding the use of animal neurobehavioral approaches in basic and pre-clinical studies. Available from https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-053.html.

PETA (2022). Pfizer bans forced swim test after PETA campaign. PETA. https://www.peta.org/media/news-releases/pfizer-bans-forced-swim-test-after-peta-campaign/

Petković, A., & Chaudhury, D. (2022). Encore: Behavioural animal models of stress, depression and mood disorders. *Frontiers in behavioral neuroscience*, *16*, 931964. https://doi.org/10.3389/fnbeh.2022.931964

Piotrowska, M. (2023). Diversity and inclusion for rodents: how animal ethics committees can help improve translation. *Journal of Medical Ethics*. doi: 10.1136/jme-2023-109166

Porsolt, R., Le Pichon, M. & Jalfre, M. (1977). Depression: a new animal model sensitive to antidepressant treatments. *Nature* 266, 730–732.

Radden, J. (2003). Is This Dame Melancholy? Equating Today's Depression and Past Melancholia. *Philosophy, Psychiatry, & Psychology, 10*(1), 37–52.

Reardon S. (2019). Depression researchers rethink popular mouse swim tests. *Nature*, *571*(7766), 456–457.

Richetto, J., Polesel, M. & Weber-Stadlbauer, U. (2019). Effects of light and dark phase testing on the investigation of behavioural paradigms in mice: relevance for behavioural neuroscience. *Pharmacology Biochemistry and Behavior* 178, 19–29.

Russell W.M.S, & Burch R.L. 1959. (as reprinted 1992). *The principles of humane experimental technique*. Wheathampstead (UK): Universities Federation for Animal Welfare.

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., ... & Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700-1712.

Schmidt, S. (2016). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Review of General Psychology 13(2): 90–100.

Shansky, R.M. (2019). Are hormones a "female problem" for animal research? *Science* 364 (6443): 825-6.

Shansky, R.M., Murphy, A.Z. (2021). Considering sex as a biological variable will require a global shift in science culture. *Nature Neuroscience* 24: 457-464.

Sorge, R.E. et al. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods* 11, 629–632.

Stanford, C. S. (2017). Confusing preclinical (predictive) drug screens with animal 'models' of psychiatric disorders, or 'disorder-like' behaviour, is undermining confidence in behavioural neuroscience. *Journal of Psychopharmacology,* 31(6), 641–643.

Sullivan, J. A. (2007). *Reliability and Validity of Experiment in the Neurobiology of Learning and Memory*. Dissertation.

Sullivan, J. A. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, *167*, 511-539.

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. *The American journal of psychiatry*, *157*(10), 1552–1562.

Trunnell, E.R., Carvalho, C. (2021). The forced swim test has poor accuracy for identifying novel antidepressants. *Drug Discovery Today* 26(12): 2898-2904.

Vialard, F., & Olivier, M. (2020). Thermoneutrality and Immunity: How Does Cold Stress Affect Disease?. *Frontiers in immunology*, *11*, 588387. https://doi.org/10.3389/fimmu.2020.588387

Voelkl, B., Altman, N. S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N. A., Kas, M. J., Schielzeth, H., Van de Casteele, T., & Würbel, H. (2020). Reproducibility of animal research in light of biological variation. *Nature reviews. Neuroscience*, *21*(7), 384–393.

von Kortzfleisch, V.T., Karp, N.A., Palme, R. *et al.* (2020). Improving reproducibility in animal research by splitting the study population into several 'mini-experiments'. *Sci Rep* **10**, 16579. https://doi.org/10.1038/s41598-020-73503-4

von Kortzfleisch, V. T., Ambrée, O., Karp, N. A., Meyer, N., Novak, J., Palme, R., Rosso, M., Touma, C., Würbel, H., Kaiser, S., Sachser, N., & Richter, S. H. (2022). Do multiple experimenters improve the reproducibility of animal studies?. *PLoS biology*, *20*(5), e3001564. https://doi.org/10.1371/journal.pbio.3001564

Wadman, M. (2023). FDA no longer needs to require animal tests before human drug trials. New law welcomed by animal welfare groups, but others say change won't happen fast. *Science*. https://www.sciencemag.org/news/2023/01/fda-no-longer-needs-require-animal-tests-human-drug-trials

World Health Organization (2021). Depression. WHO.

Wray, N.R., Ripke, S., Mattheisen, M. *et al.* (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics* 50, 668–681.

Würbel, H. (2000). Behaviour and the standardization fallacy. *Nature Genetics* 26, 263. https://doi.org/10.1038/81541

Würbel H. (2002). Behavioral phenotyping enhanced--beyond (environmental) standardization. *Genes, brain, and behavior*, *1*(1), 3–8.

Würbel, H., Chapman, R., Rutland, C. (1998). Effect of feed and environmental enrichment on development of stereotypic wire-gnawing in laboratory mice, *Applied Animal Behaviour Science* 60(1): 69-81.

Würbel, H., Stauffacher, M., Von Holst, D. (1996). Stereotypies in Laboratory Mice—Quantitative and Qualitative Description of the Ontogeny of 'Wire-gnawing' and 'Jumping' in Zur:ICR and Zur:ICR nu, *Ethology* 102(3): 371-385.

Yeoman, B. (2003). Can We Trust Research Done with Lab Mice? *Discover Magazine*, July 1, 64-71.