Validity Drifts in Psychiatric Research

Matthias Michel

Psychiatric research is in crisis because of repeated failures to discover new drugs for mental disorders. Lack of measurement validity could partly account for these failures. If researchers do not actually measure the effects of drugs on the disorders they aim to investigate, one should expect suboptimal treatment outcomes. I argue that this is the case, focusing on depression, and fear and anxiety disorders. In doing so, I show how psychiatric research illustrates a more general phenomenon that I call 'validity drift'. A validity drift occurs when, in the course of developing new measurement procedures, researchers end up studying something distinct from what they originally aimed to study. I analyse the different ways that scientists attempt to validate procedures for measuring mental disorders for the purpose of testing drugs in animal models. I show how those validation efforts might fail, thereby leading to validity drifts. Overall, this analysis highlights the complex interplay between the development of new measurement procedures, their calibration, and scientific theories of the target phenomena.

1. Introduction

Many things can go wrong when developing a new measurement procedure. The procedure might fail to reach the desired degree of accuracy. Worse, it might fail to measure what it was supposed to measure—it could be invalid. Scientists can calibrate procedures to assess and improve their accuracy (Chang [2004]; Tal [2017]). But it is not clear how validity is assessed, preserved, or lost throughout the development of new measurement procedures. This article is a case study of a domain where things went wrong (or so I argue) with respect to validity: psychiatric research. An analysis of this story of failure shows how complex the task of preserving validity really is, and reveals some of the conditions required for the development of valid measurement procedures.

Psychiatric research is in crisis. As noted by Fibiger ([2012], p. 649): 'The data are in, and it is clear that a massive experiment has failed: despite decades of research and billions of dollars invested, not a single mechanistically novel drug has reached the psychiatric market in more than 30 years'. The main drugs used today are updated versions of drugs discovered more than fifty years ago, often following serendipitous findings (Robinson [2018]; López-Muñoz et al. [2022]). Repeated failures to discover mechanistically new drugs have led the pharmaceutical industry to significantly decrease its investments in new treatments for psychiatric disorders (Miller [2010]; Fibiger [2012]; Hyman [2012], [2013]).

Lack of measurement validity could partly account for these failures. If researchers do not actually measure the effects of drugs on the disorders they aim to investigate, one should expect suboptimal treatment outcomes. I argue that this is the case, focusing on depression, and fear and anxiety disorders. In doing so, I show how psychiatric research illustrates a more general phenomenon that I call ‘validity drift’. A validity drift occurs when, in the course of developing new measurement procedures, researchers end up studying something distinct from what they originally aimed to study.

I develop the notion of validity drift in section 1, and show how validity drifts can occur when new measurement procedures are developed for the purpose of measuring a variable of interest in a new domain. In section 2, I identify several factors that contribute to validity drifts in psychiatric research. I hold that the main culprit for validity drifts is the reliance on animal models for testing drug efficacy. I describe several ways that scientists might attempt to validate indicators of mental disorders in animal models, and show how those attempts might sometimes fail, thereby leading to validity drifts. In section 3, I show how measurement procedures in humans might also be subject to validity drifts.

2. Validity and Metrological Extensions

There are two main reasons for developing new measurement procedures. The first is to increase accuracy. The second is to extend the domain of measurement—a process that Chang ([2004], p. 152) calls ‘metrological extension’. Cases of metrological extension are my main focus.

Measurement procedures have a given domain of application. Take the measurement of distance, for instance. Researchers measure distances up to 10^9 kilometres by using telemetry—send a signal, and wait for it to come back. Past that point, they need other methods. Scientists can use telemetry to calibrate the parallax method, which allows reliable measurements up to 10^3 light years. In turn, they rely on the parallax method to calibrate a method for measuring distance by relying on Cepheid variables—stars for which researchers can know the true luminosity, thereby allowing them to measure distance by comparing the true luminosity to the observed luminosity (for example, Feast and Catchpole [1997]).

This illustrates the process of metrological extension. Scientists developed new procedures to ‘extend’ the domain of measurement. Other cases of metrological extension include, for instance, the development of procedures for measuring temperature above the boiling point of mercury, or the development of procedures for measuring pressure past the point where pressure gauges break down (Chang [2004], p. 152, [2012], p. 188). For reasons that will become clear soon enough, I restrict my analysis of cases of metrological extension to cases in which we already have good reasons to believe that the old procedure is valid. The question is, then: given that one has a valid procedure for measuring a construct in a given domain, how does one develop a new, valid procedure to extend the domain of measurement?

Researchers might sometimes get lucky and develop procedures that happen to be valid for measuring the relevant construct. But they can't count on luck. Instead, they can put chances on their side by developing procedures that satisfy two conditions for successful metrological extensions. First, it should be clear that the construct of interest is present in the target domain. One can't measure something that doesn't exist (Borsboom et al. [2004]). Call this the 'existence condition'. In the case of temperature, for instance, scientists had no reason to expect that temperature would somehow stop increasing beyond the boiling point of mercury, and could thus extend the domain of measurement of temperature. While the existence condition might seem trivial for cases like temperature, pressure, or distance, it is not trivially satisfied in a wide variety of other cases, as I show in section 2.

Following Chang ([2004], p. 152), I call the second condition 'overlap': the original procedure and the new procedure should have 'an overlapping domain of application'. For example, there is an overlapping domain of application between the procedure relying on cepheid variables and the parallax method—a set of distances where the two methods apply. This case of metrological extension satisfies the overlap condition.[1] That allows scientists to calibrate the new procedure by comparing its results to those of the old procedure in the overlapping domain of application—a process that Tal ([2017]) calls 'black-box calibration'.

When a metrological extension satisfies the existence and overlap conditions, and both procedures provide the same outcomes in the domain of overlap, scientists have good—although defeasible—reasons for holding that the new procedure is valid in the new domain. That is, the procedure measures the construct of interest. In this article, I assume the definition of validity

[1] There is a small difference between how I use the phrase 'satisfying the overlap condition' and how Chang uses it. Chang ([2004], p. 152) not only requires the existence of a domain of overlap where the old and new procedures apply, but also that the two procedures provide consistent outcomes in that domain. However, for my purposes, it is important to distinguish between the existence of an overlapping domain of application, and agreement in outcomes within that domain. The way I use the phrase, satisfying the overlap condition only requires the former. The latter is a matter of successful calibration of the procedures that may or may not happen, given that the overlap condition is satisfied.

provided by Borsboom et al. ([2004], p. 1061): A procedure is 'valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure'. If a metrological extension satisfies the existence condition, the new procedure satisfies (a). If it satisfies the overlap condition and the old and new procedures provide consistent outcomes in the domain of overlap, this is (defeasible) evidence that the new procedure satisfies (b). Remember: I assumed above that the old procedure is valid. If so, the fact that the old and new procedures provide consistent outcomes in an overlapping domain is evidence that they respond equally to variations in the construct of interest. Indeed, a simple explanation for the convergence in measurement outcomes is that both procedures respond to variations in the same construct.[2] What matters for our present purposes is that satisfying the existence and overlap conditions provides good evidence that the metrological extension is successful with respect to validity: the old and new procedures measure the same construct.

Validity drifts are likely to occur when metrological extensions go wrong—when the existence and overlap conditions are not satisfied. Failing to satisfy these conditions means that scientists are not sure whether the relevant construct exists in the new domain. And there is also no domain of overlap to compare the outcomes of the old and new procedures. In the next section, I show how validity drifts can occur when the existence and overlap conditions are not satisfied. I do so with concrete cases of metrological extension: the metrological extensions required for testing the efficacy of drugs for mental disorders in animal models.

3. Failed Metrological Extension: Validity Drifts from Humans to Animals

3.1. Animal models and metrological extensions

Psychology is rife with metrological extensions. Psychologists might aim to study colour perception in infants, theory of mind in apes, or states of consciousness in non-communicating patients. All of these cases require metrological extensions. The main measurement methods used in adult humans to assess the relevant constructs are unavailable in these domains. Pretty much any case where scientists aim to detect or measure mental attributes in non-communicating animals requires developing new procedures extending the domain of measurement of

[2] What if we do not have good reasons to believe that the old procedure is valid? In that case, observing converging results in a domain of overlap does not provide evidence of validity for the new procedure. Indeed, both procedures could be invalid. As suggested by a reviewer, the fact that one obtains convergent results with, say, two different types of IQ tests, does not necessarily indicate that the tests are valid. Convergence only provides evidence of validity for one of the two tests if we have good reasons to hold that the other test is valid.

the relevant attributes. Identifying general conditions for successful metrological extensions is a pressing matter for psychology as a whole.

I focus here on the measurement of mental disorders in non-human animal models, where measurement is carried out for the purpose of testing drug efficacy. In this context, an animal model is 'a living organism used to study brain–behaviour relations under controlled conditions, with the final goal to gain insight into, and to enable predictions about, these relations in humans' (van der Staay [2006], p. 133). More specifically, given my focus on drug testing, an animal model is 'the use of a nonhuman animal, usually a mammal or vertebrate to predict human response to drugs and disease' (Shanks et al. [2009]). For instance, rodents are commonly used as animal models of depression to test anti-depressants, following various experimental manipulations (Gururajan et al. [2019]).

It should be clear that measuring the effects of drugs on mental disorders in animal models requires metrological extensions. Take the measurement of depression, for instance. Some of the main scales such as the Beck depression inventory (BDI) or the Center for Epidemiologic Studies depression scale (CES-D) are self-rating scales including items such as 'I think my life is a failure'. Even scales administered by clinicians, such as the Hamilton depression rating scale, are largely based on verbal responses provided in structured interviews. The same is true for the assessment of many other mental disorders, from obsessive compulsive disorders to eating disorders. Animals such as rodents cannot provide subjective reports about their mental states. And mental disorders are not directly observable. So, measuring the effects of drugs on mental disorders in animal models requires a metrological extension.

From the discussion in section 1, we should expect metrological extensions in this area to be particularly hard. First, as I discuss in detail in section 2.2, it is not always clear that the relevant constructs exist in animal models—metrological extensions might fail to satisfy the existence condition. Second, one might suspect that metrological extensions in psychiatric research often fail to satisfy the overlap condition. The reason we need a metrological extension in the first place is that the procedures used in humans do not apply in animal models. So, comparing the outcomes of both procedures in non-human animals is generally out of the question, although both procedures can sometimes be used in humans, as I discuss in section 3.[3] In

[3] The overlap condition does not always fail to be satisfied in animal models in psychiatric research. For instance, certain tasks allowing to measure attention deficits can be adapted in rodents (Kim et al. [2015]). This is the exception rather than the rule.

what follows, I provide examples to illustrate the various ways that metrological extensions might fail in psychiatric research, and how such failures might lead to validity drifts.

Before I do so, let me explain why this focus on animal models is relevant for understanding the current crisis in psychiatry. Two phases typically constitute drug discovery: a pre-clinical phase in which a drug is tested in non-human animals for both safety and efficacy, and a clinical phase where it is tested in humans. Attrition rates for drug development are generally high—90% on average, meaning that most drugs never end up on the market (Paul et al. [2010]; Hay et al. [2014]). Lack of efficacy in humans seems to be the main culprit (Hay et al. [2014]). Attrition rates increase significantly between pre-clinical and clinical trials—which means that pre-clinical trials involving animal models deliver a high rate of false positives (Paul et al. [2010]; Garner [2014]; Akhtar [2015]; Seyhan [2019]). As noted by Garner ([2014], p. 440), the data reveals that 'attrition is fundamentally an animal issue, not a human one'. For this reason, the metrological extensions involved in pre-clinical trials using animal models seem like a good place to start for understanding the failure to develop efficient novel drugs in psychiatry. Understanding what might go wrong at this stage is my aim here.[4]

3.2. Valid models, valid measurement, and non-existence drifts

It is important to distinguish between model validity and measurement validity.[5] Most of the discussions about the role of animal models in science—and psychiatric research in particular, focus on the validity of the models themselves (for example, Nestler and Hyman [2010]; Shelley [2010]; Belzung and Lemoine [2011]; Greek and Shanks [2011]; Sjoberg [2017]; Harro [2019]). I focus instead on the validity of measurements in animal models. While measurement and model validity are different, they are closely connected. Let me say more about this distinction.

[4] Let me emphasize again that validity drifts cannot be the whole story for the current crisis in psychiatry. Many factors contribute (see, for example, Hay et al. [2014]; Seyhan [2019]). For one thing, while lack of efficacy partly explains high attrition rates, poor safety profiles (especially when combined with other drugs) also account for many failures. Some factors are also not specific to psychiatry. For instance, many findings in biomedical research fail to replicate (Ioannidis [2016]). Perhaps more importantly, it might also be the case that for at least some disorders there are no disorder-specific neurophysiological mechanisms for the drugs to act on in the first place—no latent neurophysiological variable that constitutes the disorder. Instead, some mental disorders might be best modelled as complex, interacting psychological states (for example, Borsboom et al. [2019]). I come back to some of the consequences of this view for the validation of measurement procedures in section 2.4.

[5] This distinction is similar to the distinction developed by Maximino and van der Staay ([2019]) between models and tests.

Following Belzung and Lemoine ([2011], p. 7), 'an animal model has validity inasmuch as it is similar to a modelled human disease'. As they emphasize, different dimensions of similarity underlie different dimensions of model validity: similarity in the factors that lead to the disease (pathogenic validity); similarity in the mechanisms underlying the disease (mechanistic validity); similarity in observed behaviours (face validity); and similarity in reactions to therapeutic agents (predictive validity).[6]

What matters for our purpose is that model validity is a necessary condition for measurement validity, but not a sufficient condition. It is necessary because one can't measure something that doesn't exist. To make this salient: no experimental manipulation can make jellyfish a valid animal model for dyslexia. Because of this, all procedures for measuring dyslexia in jellyfish are invalid.

Most cases are not that obvious. Animal models are never perfect models. But no one should claim that all animal models of, say, depression, are invalid just because they do not display the complete symptomatology of depression—including, for instance, suicidal thoughts. Similarity comes in degrees. And since model validity is defined in terms of similarity, model validity comes in degrees too.

This is presumably why the '-like' locution is extremely prevalent in psychiatric research: a model can be more or less 'depression-like', or 'OCD-like'. The degree to which the model is '-like' the modelled disorder depends on its degree of similarity to the modelled disease. This degree of similarity is assessed by identifying '-like' traits: 'depression-like', 'OCD-like', or 'anxiety-like' traits. I come back to the thorny issue of what makes a trait 'disorder-like' in section 2.3. What matters for now is that if model validity is indeed a necessary condition for measurement validity, confidence in the validity of the measurement for a target disorder depends on the degree to which the model is indeed 'like' the target disorder.

While necessary for valid measurement, model validity is not sufficient. This point has not attracted as much attention. So, let me provide an example with the social defeat model of depression (Nestler and Hyman [2010]; Hollis and Kabbaj [2014]; Gururajan et al. [2019]). In the social defeat model, scientists repeatedly introduce a male rodent into the cage of an older, dominant and aggressive male, leading to an attack and subordination of the intruder. The intruder is then placed in a protective cage next to the cage of the older rodent, thereby leading to further exposure to the aggressive male. Rodents submitted to chronic social defeat display

[6] For similar distinctions, see (van der Staay [2006]).

a decreased preference for rewarding stimuli, such as a sucrose solution—often interpreted as an indicator of anhedonia (Scheggi et al. [2018]), as well as social withdrawal, reduced exploratory and locomotor activity, and reduced gains in body weight (Rygula et al. [2005]; Hollis and Kabbaj [2014]). Some of these abnormal behaviours decrease following chronic administration of (some) anti-depressants (Berton et al. [2006]; Tsankova et al. [2006]).

For the sake of argument, assume that the social defeat model is a valid model of depression: mice submitted to repeated social defeat are in a depressive-like state. This does not settle the question of determining how to measure a decrease in depressive state following the administration of a drug. To see this, suppose that one interprets an increase in exploratory behaviour as indicating a reduction in the depressive-like state. For this procedure to be valid, one not only needs a valid model; one also needs to validate the interpretation of an increase in exploratory behaviour as an indicator of a reduction in depressive state. Following Borsboom et al.'s ([2004]) account of measurement validity, this is the case if the reduction in depressive state causes the increase in exploratory behaviour, or if an increase in exploratory behaviour partly constitutes a decrease in depressive state (I develop this account in section 2.3).

Based on this analysis, one can distinguish between two main kinds of validity drifts: 'non-existence drifts' and 'mistargeting drifts'. Let me start with the former and illustrate with an example. I discuss the latter in the next section.

Non-existence drifts are cases in which scientists believe that they are studying a construct in an animal model that does not in fact instantiate the relevant construct. This kind of validity drift exemplifies a failure to satisfy the existence condition for successful metrological extension. Non-existence drifts in psychiatric research are well illustrated in some cases using the 'forced-swim test' (FST) for evaluating the effects of drugs on depression in rodents. The FST is commonly used to assess depression-like states in mice.[7] Experimenters place a mouse in a cylinder filled with water and record its behaviour. During an initial phase, called the 'active' phase, the mouse tries to escape by swimming, or climbing. As the test progresses, it exhibits a period of immobility—the 'passive' phase. Experimenters rescue the animals after a fixed amount of time (usually between five and ten minutes). This is often followed by a second test

[7] Molendijk and de Kloet ([2015]) estimated that as of 2015 about two thousand studies had used the forced swim test to evaluate the anti-depressant effects of drugs. A more recent survey by Molendijk and de Kloet ([2019], p. 6) found that the number of studies using this test is still growing, and that 'the majority of the papers (about 72%) qualifies the behaviour of the floating rodent as depressive-like'.

(FST2), which repeats this manipulation after the administration of a drug. The typical interpretation is that immobility is an indicator of 'despair'—a depression-like trait (Prosolt et al. [1977]). The capacity of a drug to delay the passive phase is thus interpreted as a measure of its efficacy as an anti-depressant.

While the FST is still widely used, most researchers evaluating its validity now consider it an invalid animal model (Nestler and Hyman [2010]; Molendijk and de Kloet [2015], [2019]; Commons et al. [2017]; Gururajan et al. [2019]; Armario [2021]). Put simply, the test does not induce a depression-like state. An alternative interpretation is that immobility in the forced-swim test is an adaptive strategy: floating saves energy (Hawkins et al. [1978]). Nishimura et al. ([1987], p. 94) noted early on that rodents who entered the passive phase more quickly had a lower chance of sinking within two hours, concluding that 'immobility is beneficial in preventing the rats from sinking'. According to this interpretation, the transition to immobility reflects behavioural adaptation to an acute stressor (Molendijk and de Kloet [2015]). There is nothing 'depression-like' (or pathological) in selecting the strategy that gives one the best chance of making it out alive.[8]

Studying depression with the FST is a plausible case of validity drift—a non-existence drift. Researchers might believe that they are measuring the effects of drugs on a construct—depression, or a 'depressive-like' state—that the target model does not in fact instantiate. If the FST does not induce a depressive-like state, all measurement procedures purporting to measure the effects of drugs on that state are invalid. What immobility in the FST really indicates is currently unclear. Instead of the anti-depressant effects of drugs, scientists might have been measuring their effects on other cognitive processes, such as learning, memory consolidation, motivation, the switch to an active coping strategy, or a combination of all of those and other unknown factors (Molendijk and de Kloet [2015]; Commons et al. [2017]; Armario [2021]). Several other closely related tests in animal models of depression might exemplify the same kind of non-existence drift, such as the tail suspension test (Steru et al. [1985]), and the zebrafish tail immobilization test used in zebrafish models of depression (Lachowicz et al. [2021]).[9]

[8] You might ask: why do anti-depressants show an effect on the amount of time before the immobility response is triggered? Molendijk and de Kloet's ([2015]) hypothesis is that anti-depressants interfere with learning and storing the immobility response in memory. They write: 'the driving force behind the observed switch to immobility […] in the FST is not "depression" or "despair" but the gift of learning and memory, which promotes behavioural adaptation and survival' ([2015], p. 390).

[9] For a critical assessment of zebrafish models of depression, see (de Abreu et al. [2018]).

Non-existence drifts occur because invalid models can be expected to lead to invalid measurement. It should be clear that these validity drifts come from failing to satisfy the existence condition for metrological extensions. In the remainder of this paper, I assume, for the sake of argument, that the animal models I consider have at least sufficient validity to prevent non-existence drifts. My purpose is to illustrate a more complex kind of validity drift: mistargeting drifts—cases where the target system does instantiate the construct of interest but researchers fail to develop valid procedures for measuring it.

3.3. Validating measures of disorders: The problem of coordination

A measurement outcome results from interpreting an indicator (or set of indicators) as indicating a given value for the relevant construct. For instance, one interprets the height of the column of mercury in a thermometer—an indicator—as indicating a certain temperature value. In psychiatry, researchers typically measure the effects of drugs on disorders by interpreting variations in certain ‘disorder-like’ traits as indicating variations in the relevant construct. Disorder-like traits are indicators. Researchers interpret them as indicating the presence of disorder-like states. For instance, in animal models, low locomotor activity is an indicator that might be interpreted as indicating a depression-like state (for example, Gururajan et al. [2019]).

When scientists aim to extend the domain of measurement of mental disorders to assess disorder-like states in animal models, they need to find indicators that can be (validly) interpreted as indicating the relevant disorders. Non-existence drifts occur when the model does not instantiate the relevant disorder-like state. I set non-existence drifts aside from now on. Mistargeting drifts occur when interpreting a given indicator as indicating a disorder is not valid. This is the case if the disorder does not cause variations in the observed indicator, and yet researchers interpret those variations as indicating variations in the disorder. As a result, the measurement procedure misses its target—the measurement outcome does not actually reflect variations in the disorder. The main questions I now aim to answer are: What makes it the case that interpreting a ‘disorder-like’ trait as indicating a given disorder is valid? And how can scientists limit mistargeting drifts? How can they find out whether a given disorder-like trait is a valid indicator for a given disorder?

‘Disorder-like’ traits come in two main varieties: behavioural traits and unobservable, non-behavioural traits. For instance, in the social defeat model mentioned above, a reduction in exploratory behaviour in rodents is interpreted as a behavioural ‘depression-like’ trait, and variations in that trait are interpreted as indicating variations in a depressive-like state (Hollis and

Kabbaj [2014]; Gururajan et al. [2019]). Variations in unobservable traits, such as disruptions to mechanisms hypothesized as relevant for the disorder, are also sometimes interpreted as indicators of the disorder. For instance, in the social defeat model, a dysregulation of the stress responsive hypothalamic pituitary adrenal (HPA) axis is interpreted as indicating a depressive-like state (for example, Hollis and Kabbaj [2014]).

In order to identify conditions required for a disorder-like trait to count as a valid indicator of the relevant disorder,[10] one needs to understand what makes a trait 'disorder-like' in the first place. While the '-like' locution is extremely common in psychiatric research, to the best of my knowledge, there is no systematic analysis of what makes a trait 'disorder-like'.[11]

My tentative account is the following: Where *x* is a mental disorder, a trait is *x*-like, if, and only if, the trait is directly caused by *x* (or an *x*-like state); or the trait (at least partly) constitutes *x* (or an *x*-like state). According to this account, what makes a trait such as a reduction in exploratory behaviour 'depression-like' is that it is caused by a depression-like state (or partly constitutes it). Similarly, a dysregulation of the HPA axis could be called 'depression-like' if it partly constitutes a depressive-like state in the target model.

Given this definition, it should be clear that interpreting variations in a trait as indicating variations in a target disorder is valid if, and only if, the trait is disorder-like. Indeed, if the trait is disorder-like, it is either caused by, or partly constitutes, the relevant disorder. This guarantees that the relevant disorder-like state exists, and that variations in the disorder-like state produce variations in the outcome of the procedure. This, in turn, satisfies our definition of measurement validity (Borsboom et al. [2004]). In sum, validating the interpretation of a trait as an indicator of a disorder-like state amounts to confirming that the relevant trait is indeed disorder-like. This kind of validation could prevent validity drifts, and thus enable successful metrological extensions. My goal is to understand how scientists can do this.

[10] I sometimes talk about 'valid indicators'. This is just a shorter way of saying that interpreting a given indicator as indicating a given construct is valid. In and of themselves, indicators are neither valid nor invalid. But one can have valid or invalid interpretations of what those indicators indicate.

[11] This lack of clarity is probably what prompts researchers, like Garner ([2014], p. 443), to dismiss the notion: 'The word "-like" (as in "OCD-like" or "anxiety-like") has become pervasive in behavioural neuroscience, but it represents an incredibly dangerous slip in logic. The trap is simple to understand: calling a measure "-like" does not make it so (that is an empirical issue of validity). Calling a measure "-like" is a rhetorical device that gives the measure a sheen of respectability and scientific caution, while in truth masking the fact that no attempt has been made to validate the measure or that it is being used despite being known to be invalid. In either case, this is simply bad science'. Contrary to Garner, I do not believe that we should dismiss the notion entirely. The '-like' locution is useful to emphasize that models are never complete models: model validity comes in degrees.

At this point researchers face a significant threat of circularity. Validating measurement procedures requires making sure that the traits interpreted as indicators of the target disorder are indeed 'disorder-like'. To make sure that a trait is disorder-like, we need reasons to hold that it is caused by or partly constitutes the disorder. However, in turn, determining whether or not a trait is caused by (or partly constitutes) the disorder would seem to require valid measurement of the disorder. So, validating measurement procedures seems to require valid measurement.

I take this threat of circularity to be an instance of the 'problem of coordination', which applies to a wide variety of attempts to develop measurement procedures in domains in which one does not already have valid procedures (Chang [2004]; van Faassen [2008]; Barwich and Chang [2015]; Padovani [2015]; Tal [2020]; Michel [2023]). The problem is the following: validating a measurement procedure requires identifying indicators that co-vary with the relevant construct. But determining whether any given indicator co-varies with the relevant construct seems to require valid measurement of that construct. So, validating measurement procedures seems to require valid measurement.

There is no simple recipe to solve the problem of coordination. It has to be solved on a case-by-case basis. In psychiatric research, I identify three main ways that scientists have attempted to validate measurement procedures. First, identify indicators that vary following the administration of a drug known to have effects on the relevant disorder in humans—call this 'validation by drug responsiveness'. Second, identify indicators that systematically co-vary with several other indicators that are face-valid indicators of the relevant disorder—call this 'validation by co-variation'. Third, rely on a theory of the relevant disorder that hypothesizes that the indicator is an effect of (or partly constitutes) the disorder—call this 'validation by theory'. I present the first two validation methods in the next section and explain how they might sometimes fail to deliver on their promises, thereby leading to validity drifts. I leave my discussion of validation by theory for section 3.

3.4. Validation by drug-responsiveness and co-variation

Scientists can attempt to validate an indicator by analysing how it responds to the administration of a drug that is known to have effects on the relevant disorder in humans. Validity drifts are likely following this process of 'validation by drug responsiveness'. This is because nothing guarantees that the observed effect of the drug on the indicator results from a reduction in the relevant disorder-like state.

Let me illustrate: As explained above, the fact that immobility in the forced-swim test decreases following the administration of anti-depressants was interpreted by Prosolt et al. ([1977]) as validating immobility as an indicator of a depressive-like state in mice. The problem with this validation method is that it assumes that all observable effects of anti-depressants result from their anti-depressant effects. In the FST, anti-depressants reduce immobility, but not in virtue of their anti-depressant effects—that is, not in virtue of their effects on a depressive-like state. Drugs for mental disorders have many behavioural effects that are independent of their effects on the target disorders. So, showing that an indicator varies following the administration of a drug falls short of validating that indicator, since nothing guarantees that the effect of the drug on the indicator goes through its effect on the relevant disorder.[12] Apparent validation by drug-responsiveness could in turn cause validity drifts by leading researchers to believe that an indicator is valid when it is not.

A more promising validation method to avoid validity drifts is validation by co-variation. It consists in analysing co-variations between an indicator and a set of other indicators that are also face valid indicators for the target disorder. Let me explain.

Indicators of psychiatric disorders are not solitary creatures: they often co-vary with other effects or constitutive parts of the relevant disorder. For instance, as noted above, the social defeat model of depression exhibits a decreased preference for a sucrose solution, social withdrawal, reduced exploratory and locomotor activity, and reduced body weight (Hollis and Kabbaj [2014]). The fact that an indicator like locomotor activity varies at the same time as all these other indicators might count as (defeasible) evidence that it is a valid indicator of depression. Indeed, a simple explanation for this systematic co-occurrence is that all those indicators have a common cause: a depressive-like state. Since positing a common cause provides a plausible explanation for the co-variation of indicators, co-variation provides evidence of validity. But since a common cause is not the only explanation, this evidence is defeasible. Co-variation alone, therefore, does not entirely prevent validity drifts. In turn, evidence of co-variation can be complemented by the administration of a drug known to have anti-depressant effects in humans and by showing a common effect of the drug on all these indicators. I call this way of validating measurement procedures ‘validation by co-variation’.

[12] Aside from the risk of validity drifts, this validation method also likely leads to false negatives in the long run. Indicators are interpreted as valid mostly because they vary following the administration of known anti-depressants. Mechanistically novel drugs that are tested but do not show an effect on those indicators are then deemed ineffective. However, mechanistically novel drugs might be effective at reducing aspects of depression that are not reflected by these indicators.

Importantly, scientists can also successfully apply validation by co-variation to demonstrate a lack of validity. For instance, Mul et al. ([2016]) showed that mice exposed to repeated forced swimming over five days showed a progressive increase in immobility in the FST. However, this increase in immobility did not co-occur with a change in other indicators also commonly interpreted as indicating a depressive-like state: decreased sucrose preference, body weight, locomotor activity, and exploratory behaviour. Had repeated forced swimming caused an increase in immobility through an increase in a depressive-like state, we should have observed a concomitant increase in other (purportedly) depression-like traits. This is not what Mul et al. ([2016]) found. They concluded that variations in immobility in the FST are not caused by variations in a depressive-like state.

While promising, validation by co-variation comes with two main drawbacks, which might lead to validity drifts. First, the method requires several co-varying indicators that have a reasonable chance of being valid indicators for the relevant disorder. This is not always the case. Indeed, symptoms are often shared by multiple disorders, some of which might also have high rates of comorbidity. For example, the DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders*; American Psychiatric Association [2013]) distinguishes generalized anxiety disorders from major depressive disorders, but the two disorders have a high degree of comorbidity and many symptoms in common. These different disorders might therefore share a set of co-varying indicators. For this reason, it is unclear whether a systematic co-variation of indicators should count as evidence that each indicator is valid for a target disorder, instead of a closely connected disorder. The co-variation of indicators could be explained by the occurrence of several disorders with a high rate of comorbidity. For instance, comorbidity between generalized anxiety disorders and major depressive disorders might give rise to systematically co-varying indicators, only some of which can be validly interpreted as indicating a depressive-like state, while the other indicators co-vary with an anxiety-like state. This risk is especially present in the study of depression given that depression-like states in animal models are often induced by stressful situations. In those conditions, determining whether indicators co-vary with stress or a depression-like state is challenging. Apparent validation by co-variation could thus lead to mistargeting drifts. A researcher might, for instance, misinterpret an indicator of stress as an indicator of depression.

The second problem with validation by co-variation is that its generalizability is severely limited. The reason for this is that whether or not a trait is ‘disorder-like’ is highly context dependent. There is nothing intrinsically ‘depression-like’ in low exploratory activity. But it can be depression-like in a given context if it is caused by a depressive state, or partly constitutes a depressive state. The problem is that the context of validation, in which the indicator is shown to co-vary with other indicators, and the context in which scientists subsequently interpret the indicator, might differ quite significantly. An indicator validated by co-variation in a given context might be invalid in a different context, since nothing guarantees that the relevant indicator is still caused by (or partly constitutes) the relevant construct across contexts.

Let me illustrate: One might hold that reduced exploratory activity in rodents is a valid indicator of depression as a result of an experiment showing that it co-varies with many other indicators of depression. Outside the context of validation, however, low exploratory activity might have different causes. For instance, in a different context, a rodent might reduce its exploratory activity because it fears the potential presence of a predator. In that context, interpreting low exploratory activity as indicating depression is invalid.

This context-dependence is especially important for disorders that are best modelled using the 'network approach' to psychopathology (Borsboom and Cramer [2013]; Fried and Cramer [2017]; Borsboom et al. [2019]; Robinaugh et al. [2020]). The network approach conceives of mental disorders as complex networks of interacting symptoms rather than latent entities. This perspective rejects the traditional model where symptoms are manifestations of an underlying disorder, proposing instead that symptoms interact and influence each other, creating a network. For instance, insomnia may lead to fatigue, which exacerbates concentration problems, in turn inducing worry, and so on, eventually leading to a self-sustaining network of symptoms that constitutes a disorder. Following this approach, what makes a trait such as low exploratory activity depression-like is the fact that it is embedded in a network of symptoms—hence my insistence on saying that symptoms might partly constitute disorders instead of being caused by disorders. Taken out of this network, there is nothing 'depression-like' about low exploratory activity. Whether or not this trait partly constitutes a depression-like state thus depends on the context in which it is embedded. Given this, it cannot be interpreted as an indicator of depression in isolation, irrespective of its co-occurrence with other traits.

Because validity is context-dependent in this way, one cannot assume that the co-variation of a variety of indicators in a model (for instance, the social defeat model) constitutes good evidence for the validity of a trait (for example, decreased locomotor activity) as an isolated indicator of a depressive-like state across contexts. At best, it could be considered valid in the context in which the co-variation has been observed. Outside of this context, however, it needs to be re-validated to make sure that it is still caused by (or partly constitutes) a depression-like state. This can be done by verifying that the indicator still co-varies with other relevant indicators in the new experimental setting.

However, this leads us to a third problem with validation by co-variation: scientists often consider it impractical. Validation by co-variation requires a battery of tests instead of a single test. As noted by Kalueff et al. ([2008]) and Maximino and van Der Staay ([2019]), many of the tests used to evaluate the effects of drugs only output effects on a few indicators (for instance, immobility in the FST), and cannot be used to evaluate effects on a wide variety of disorder-like traits. Validation by co-variation therefore requires multiple cohorts of animals, longer testing time, and extensive laboratory resources. Furthermore, as Maximino and van Der Staay ([2019], p. 7) write: ‘the use of multiple dependent variables—either using hybrid test conditions or test batteries—dramatically reduces power, due to the requirement to use corrections for multiple comparisons, decreasing reliability and, as a consequence, increasing the number of animals needed for discovery’. For these reasons, while validation by co-variation does not generalize well across contexts, scientists often do not re-validate indicators in their new experimental settings. As such, validation by co-variation opens the door for a significant risk of validity drift: scientists might interpret isolated indicators that were previously validated in a different context as if they were still valid in the new experimental setting.

4. Validity and Theory

The final form of validation to consider is ‘validation by theory’. A theory might provide good reasons for holding that variations in measurement indicators are caused by variations in the relevant construct—good reasons for believing that the relevant procedure is valid.[13] This is what I call ‘validation by theory’.

I now present an example of validity drift caused by the reliance on a (potentially) incorrect theory, using psychiatric research on fear and anxiety disorders as a case study. This case further demonstrates how validity drifts manifest in human psychopathology research when indicators initially applied in animal models are imported back to the human case

4.1. The search for accurate measurement of fear and anxiety

[13] By ‘theory’ here I simply mean a set of hypotheses describing how the target construct relates to the relevant indicators. Those theories might sometimes be naïve theories, or implicit theories—just as one might interpret smiling as an indicator of happiness as part of a naïve theory of mind.

Fear and anxiety disorders include disorders such as phobias, social anxiety disorder, panic disorder, and panic attacks.[14] The main indicators interpreted as indicating fear and anxiety in animal models—mostly rodents—are behavioural and physiological indicators (Taschereau-Dumouchel et al. [2022]). Behavioural indicators include freezing, flight, avoidance, and exploratory behaviour in the open field and elevated plus maze tests (Steimer [2011]). Physiological indicators include skin conductance and heart rate. I call these indicators 'bio-behavioural' indicators.

Procedures relying on bio-behavioural indicators—'bio-behavioural procedures'—have not been developed solely for the purpose of a metrological extension in animal models. Instead, these procedures have also been repurposed to increase measurement accuracy in humans. Scepticism about the accuracy of subjective reports in this area of psychiatric research drives the search for alternative indicators. As noted by Fanselow and Pennington ([2018], p. 3): 'the reason the field moved away from subjective report is no mystery: they are often difficult to reliably quantify and subject to diverse response biases that can variably over-/under-estimate the subjective experience of fear. The demand characteristics of the situation may also influence self-report: for example, fear may be under-reported by a dedicated soldier and over-reported by someone wishing to persuade a physician to prescribe medications'.[15] In line with this scepticism about the accuracy of subjective reports, developing 'objective' measurement procedures for mental disorders based on observable behaviour and neurobiological indicators

[14] Fear and anxiety are often studied together. As defined in the DSM (American Psychiatric Association [2013]), fear is a reaction to immediate danger, whereas anxiety is a reaction to an anticipated (potentially unspecified) threat. Post-traumatic stress disorders (PTSD) used to be categorized as fear and anxiety disorders but are now categorized as a trauma- and stressor-related disorder in DSM-5. This decision was criticized because fear is a central component for PTSD (Zoellner et al. [2011]).

[15] Fanselow and Pennington ([2018], p. 9) further note the fruitfulness of bio-behavioural measures in the history of research on fear and anxiety: 'Perhaps the greatest leap forward in the treatment of anxiety disorders was Wolpe's extinction/exposure oriented approach […] which provided the basis for modern cognitive/behavioural therapy. It is worth remembering that Wolpe based his treatment regimen entirely on Ivan Pavlov's and Clark Hull's behavioural observations of dogs and rats, a testament to the clinical utility of bio-behavioural metrics in the treatment of fear'. The reference to (Wolpe [1958]) is somewhat ironic. Indeed, Wolpe himself did not primarily use bio-behavioural metrics to assess fear and anxiety in humans. Instead, he created the subjective units of distress scale for subjective reports of anxiety (Wolpe [1969]).

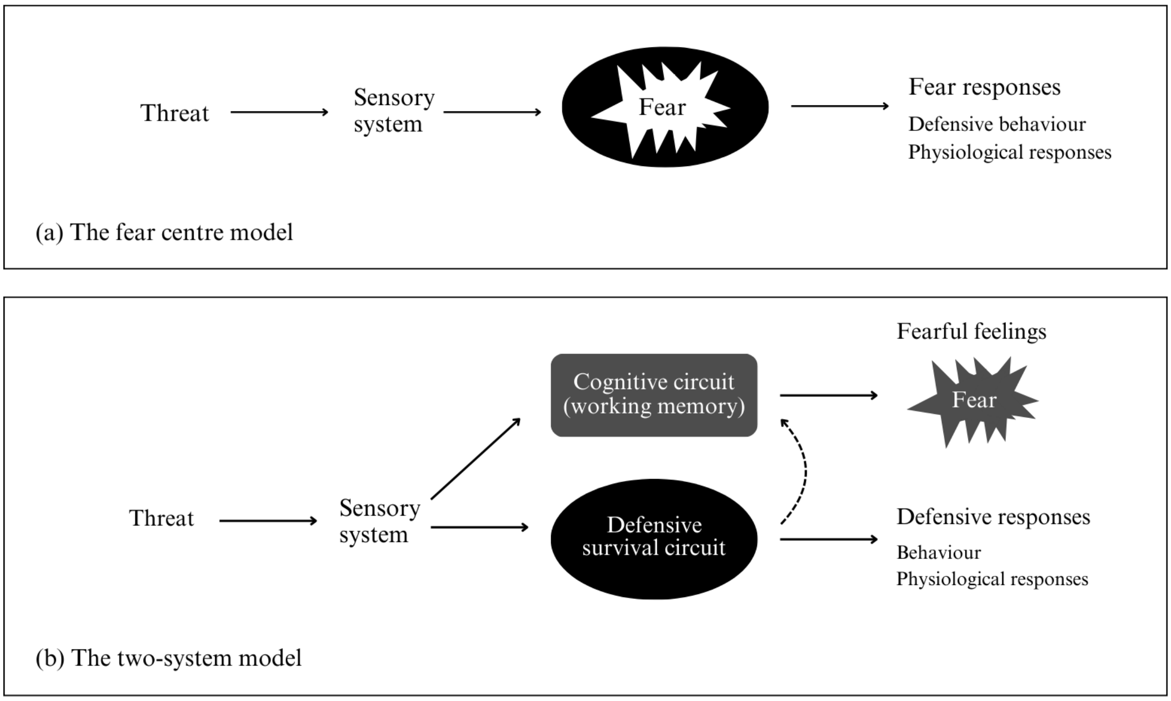

Figure 1. Two models of fear: (a) The fear centre model. Activity of a subcortical circuit is sufficient for the conscious experience of fear, and this circuit also drives bio-behavioural responses. (b) The two-systems model. Subcortical mechanisms operate un-consciously and drive bio-behavioural responses to threats. The conscious feeling of fear is generated by a separate, cognitive circuit. Source: (LeDoux and Pine [2016]).

is one of the central pillars of the research domain criteria (RDoC) initiative by the National Institute of Mental Health in the US (Kozak and Cuthbert [2016]; Morris et al. [2022]). In opposition to the central diagnostic role of subjective reports found in the DSM, the RDoC initiative has relatively marginalized self-reports in favour of behavioural and biomarkers of psychopathology (Taschereau-Dumouchel et al. [2022]). As such, researchers not only rely on behavioural and physiological indicators for the purposes of metrological extension in animal models, they also hope to increase measurement accuracy in humans.

I now argue that relying on bio-behavioural procedures to measure fear and anxiety depends on an implicit commitment to a ‘fear centre model’ of fear. If this model is incorrect, a validity drift might have occurred in fear and anxiety research.

4.2. Two models of fear and anxiety

Two of the main models of fear and anxiety are the ‘fear centre model’ (for example, Fendt and Fanselow [1999]; Adolphs [2013]), and the ‘two systems model’ (LeDoux and Pine

[2016]) (see fig. 1). These models have different implications for the validity of bio-behavioural procedures (Schaffner [2020a], [2020b]). According to the fear centre model, behavioural indicators, physiological indicators, and subjective reports, all result from the activity of a single subcortical 'fear centre', in which the amygdala plays a central role (Fendt and Fanselow [1999]; Lang and Davis [2006]; Panksepp et al. [2011]; Adolphs [2013]; Fanselow and Pennington [2018]). Neuronal activity in this fear centre gives rise to the subjective feeling of fear (Panksepp et al. [2011]). I freeze, report feeling afraid, and my heart rate increases, because I feel afraid.

Contrast this with the two systems model (LeDoux [2012], [2015]; LeDoux and Pine [2016]; LeDoux and Brown [2017]; Taschereau-Dumouchel et al. [2022]). This model distinguishes two systems: a defensive survival circuit, and a cognitive circuit. The former is involved in fast, automatic, unconscious responses to threats. Its activity causes physiological and basic behavioural responses. It is an evolutionarily ancient subcortical mechanism (LeDoux [2019]). Meanwhile, a cognitive circuit generates the conscious experience of fear and is involved in the production of subjective reports. This mechanism depends on cortical activity and appeared more recently in evolutionary history. Following the two systems model, bio-behavioural indicators and subjective reports are typically triggered by similar distal causes: the presence of a threat. However, the two systems also operate (partly) independently. The proximal causes of subjective reports and bio-behavioural indicators differ. The unconscious threat detection system triggers bio-behavioural responses. The feeling of fear causes subjective reports.

If the fear centre model is correct, interpreting variations in bio-behavioural indicators as indicating variations in fear is valid, since fear causes those variations. If the two systems model is correct, interpreting variations in bio-behavioural indicators as indicating variations in fear is not necessarily valid, since the subjective feeling of fear does not necessarily cause those variations (this would constitute a mistargeting drift).[16] Assessing which model is correct is therefore crucial with respect to measurement validity (Schaffner [2020]; Taschereau-Dumouchel et al. [2022]).

The existence of cases in which bio-behavioural indicators and subjective reports dissociate—called 'discordance' cases—seems to support the two systems model (Rachman and

[16] Following the two-systems model, it is also not clear whether animals such as rodents do in fact instantiate the subjective feeling of fear. Although the model itself is neutral on this, it might turn out that they do not. If this is the case, a non-existence drift would have occurred not only in psychiatric research using rodent models, but also more generally in the scientific research investigating the mechanisms of fear in rodents.

Hodgson [1974]; Taschereau-Dumouchel et al. [2022]). For instance, researchers have shown that subliminal presentations of threatful stimuli can increase amygdala activity and trigger physiological responses to threats, without any awareness of the stimuli, or any reported feeling of fear (for example, Hamm et al. [2003]; Tamietto et al. [2009]; Lapate et al. [2016]; Taschereau-Dumouchel et al. [2018]; for a review, see Tamietto and De Gelder [2010]). More generally, concordance—cases in which bio-behavioural indicators and subjective reports closely co-vary—has been difficult to come by. As noted by Hollenstein and Lanteigne ([2013], p. 1) in a review of the literature: 'Concordance has been weakly supported by the data, at best, but often not supported, with some research even showing evidence for the opposite, *discordance* or negative associations'. The two systems model provides an intuitive explanation for discordance cases: bio-behavioural indicators and subjective reports dissociate because they result from the activity of different, independent neuro-cognitive systems.

However, proponents of the fear centre model can also account for discordance cases. As noted by Fanselow and Pennington ([2018], p. 7) 'Despite emanating from a central generator, components of the fear response are undoubtedly born of distinct effectors, capable of being independently modulated'. A single fear centre is compatible with partly independent response systems. Consider a case of discordance in which subliminally presented stimuli elicit physiological responses in the absence of fear reports. Proponents of the fear centre model could account for this case by holding that while the subject did experience fear, as indicated by the physiological response, the system responsible for subjective reports was not sensitive enough to output a report. This view amounts to holding that subjective reports are inaccurate.

While this account might seem *ad hoc*, proponents of the fear centre model have principled reasons for believing that subjective reports are less accurate than bio-behavioural procedures. The causal path from the fear centre to behavioural and physiological indicators is more direct than that between the fear centre and subjective reports—which require the involvement of 'higher-level', cortical areas. As Fanselow and Pennington ([2018], p. 8) argue, 'the additional machinery needed to generate subjective report probably adds additional noise, rendering it, as many previous to us have suggested, a less pure and objective measure of fear'. Therefore, a fear centre model suggests that bio-behavioural procedures are more accurate than those relying on subjective reports. In that case, cases of discordance can be accounted for by the poor accuracy of subjective reports.

The two-systems model and the fear centre model can both account for cases of discordance. According to the former, discordance occurs because subjective reports are valid indicators, while bio-behavioural procedures are invalid. According to the latter, discordance occurs because bio-behavioural procedures are more accurate than subjective reports. This is a concrete case of what Tal ([2019]) calls 'the problem of quantity individuation'. Suppose that two measurement procedures, which purport to measure the same construct, provide different measurement outcomes. Two responses are available. One of the procedures (or both) could be inaccurate. Or the procedures could measure different constructs. In other words, either at least one of the procedures is inaccurate, or at least one of them is invalid. The fear centre model holds that discordance is explained by differences in accuracy. The two systems model holds that discordance is explained by differences in validity.

My goal is not to arbitrate between those two views. Instead, I use this case to highlight the interdependence between measurement validity and theory, and to show how the reliance on a potentially incorrect theory could lead to a validity drift. Relying on bio-behavioural procedures for measuring fear and anxiety commits one to a fear centre model—a model in which fear causes the relevant indicators. The hope of increasing measurement accuracy by using bio-behavioural procedures depends on a (potentially implicit) commitment to the fear centre model. However, while intuitive and parsimonious, this model might be wrong. It is, therefore, extremely important to make this (potentially implicit) commitment to a fear centre model explicit. If the two systems model is correct, bio-behavioural indicators could vary irrespective of variations in fear and anxiety. And bio-behavioural responses could even occur in non-human animals incapable of consciously experiencing fear and anxiety in the first place.

A validity drift might have occurred because of the researchers' reliance on an intuitive theory of fear: the fear centre model. By relying on bio-behavioural measurement procedures, researchers might have been studying unconscious threat detection, when they thought they were investigating fear and anxiety. The target of measurement drifted from fear and anxiety, to unconscious threat detection.

Fear and anxiety research illustrates that the link between measurement indicators and psychological constructs might often rest on implicit theoretical assumptions. Given that the relevant theoretical assumptions are often uncertain in psychology—we don't know all that much about the mind after all, this likely constitutes a significant source of validity drifts when studying other psychological constructs as well.

5. Conclusion and Future Directions

I introduced and illustrated the notion of validity drifts, focusing on the case of psychiatric research on depression, and fear and anxiety. As they develop new measurement procedures, either to extend the domain or measurement or to increase accuracy, researchers sometimes end up measuring something distinct from what they originally aimed to measure.

I identified several reasons for validity drifts in psychiatric research. First, the target construct might simply not be present in the relevant domain, as in animal models. Second, validation efforts might have failed while providing an illusion of validity. For instance, we saw that drug-responsiveness is not sufficient to establish the validity of an indicator, nor is co-variation with other indicators. Finally, validity drifts can occur when researchers deem a procedure valid by relying on incorrect theories.

All in all, the view according to which validity drifts occurred in psychiatric research could be part of the explanation for the repeated failures to identify mechanistically novel drugs for mental disorders (Miller [2010]; Fibiger [2012]; Hyman [2012], [2013]). If researchers fail to evaluate the effects of drugs on the constructs human patients care about—and evaluate the effects on some other construct instead, one should expect suboptimal treatment outcomes.

My goal was to shed light on the issue of validity drifts in psychiatric research. I did not say much about potential solutions. Let me finish by suggesting three ways that research could potentially mitigate validity drifts.

First, since we saw that successful metrological extensions require a domain of overlap, developing measurement procedures that can be used in both human and non-human animals might help. For instance, as noted above, some procedures used in humans for measuring attention deficits can be adapted in rodents (Kim et al. [2015]). Second, I argued that despite some shortcomings, a promising method for validating measurement procedures is validation by co-variation—combined with validation by drug-responsiveness. Studies explicitly dedicated to using these validation methods for validating measurement procedures should be encouraged (for example, Mul et al. [2016]). Finally, I argued that validity drifts sometimes occur when relying on incorrect theories. What Tal ([2017]) called ‘white-box’ calibration (or ‘model-based’ calibration; Michel [2019]) could be the answer. Researchers can calibrate a measurement procedure by developing a model of the way it interacts with the relevant construct. As we saw in the case of the measurement of fear and anxiety, determining which model of fear is correct would go a long way towards avoiding validity drifts.

Acknowledgments

I thank two anonymous reviewers, Anouk Barberousse, Ned Block, Axel Constant, Joseph LeDoux, Edouard Machery, Luke Roelofs, and Eran Tal for helpful discussions and comments.

Center for Mind, Brain and Consciousness
New York University
New York, USA
matthias.michel.curtil@gmail.com

References

Adolphs, R. [2013]: 'The Biology of Fear', *Current Biology*, **23**, pp. 79–93.

Akhtar, A. [2015]: 'The Flaws and Human Harms of Animal Experimentation', *Cambridge Quarterly of Healthcare Ethics*, **24**, pp. 407–19.

American Psychiatric Association [2013]: *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, Washington, DC: American Psychiatric Association Publishing.

Barwich, A. S. and Chang, H. [2015]: 'Sensory Measurements: Coordination and Standardization', *Biological Theory*, **10**, pp. 200–11.

Belzung, C. and Lemoine, M. [2011]: 'Criteria of Validity for Animal Models of Psychiatric Disorders: Focus on Anxiety Disorders and Depression', *Biology of Mood and Anxiety Disorders*, **1**, available at <doi.org/10.1186/2045-5380-1-9>.

Benke, C., Krause, E., Hamm, A. O. and Pané-Farré, C. A. [2018]: 'Dynamics of Defensive Response mobilization during Repeated Terminations of Exposure to Increasing Interoceptive threat', *International Journal of Psychophysiology*, **131**, pp. 44–56.

Berton, O., McClung, C. A., DiLeone, R. J., Krishnan, V., Renthal, W., Russo, S. J., Graham, D. et al. [2006]: 'Essential Role of BDNF in the Mesolimbic Dopamine Pathway in Social Defeat Stress', *Science*, **311**, pp. 864–68.

Borsboom, D., Mellenbergh, G. J. and Van Heerden, J. [2004]: 'The Concept of Validity', *Psychological Review*, **111**, pp. 1061–71.

Borsboom, D. and Cramer, A. O. [2013]: 'Network Analysis: An Integrative Approach to the Structure of Psychopathology', *Annual Review of Clinical Psychology*, **9**, pp. 91–121.

Borsboom, D., Cramer, A. O. J. and Kalis, A. [2019]: 'Brain Disorders? Not Really: Why Network Structures Block Reductionism in Psychopathology Research', *Behavioral and Brain Sciences*, **42**, available at <doi.org/10.1017/S0140525X17002266>.

Bulteel, K., Ceulemans, E., Thompson, R. J., Waugh, C. E., Gotlib, I. H., Tuerlinckx, F. and Kuppens, P. [2014]: 'DeCon: A Tool to Detect Emotional Concordance in Multivariate Time Series Data of Emotional Responding', *Biological Psychology*, **98**, pp. 29–42.

Chang, H. [2004]: *Inventing Temperature: Measurement and Scientific Progress*, Oxford: Oxford University Press.

Chang, H. [2012]: *Is Water H_2O? Evidence, Realism, and Pluralism*, Dordrecht: Springer.

Commons, K. G., Cholanians, A. B., Babb, J. A. and Ehlinger, D. G. [2017]: ‘The Rodent Forced Swim Test Measures Stress-Coping Strategy, Not Depression-Like Behavior’, *ACS Chemical Neuroscience*, **8**, pp. 955–60.

de Abreu, M. S., Friend, A. J., Demin, K. A., Amstislavskaya, T. G., Bao, W. and Kalueff, A. V. [2018]: ‘Zebrafish Models: Do We Have Valid Paradigms for Depression?’, *Journal of Pharmacological and Toxicological Methods*, **94**, pp. 16–22.

Fanselow, M. S. and Pennington, Z. T. [2018]: ‘A Return to the Psychiatric Dark Ages with a Two-System Framework for Fear’, *Behaviour Research and Therapy*, **100**, pp. 24–29.

Feast, M. W. and Catchpole, R. M. [1997]: ‘The Cepheid Period-Luminosity Zero-Point from *Hipparcos* Trigonometrical Parallaxes’, *Monthly Notices of the Royal Astronomical Society*, **286**, available at <doi.org/10.1093/mnras/286.1.L1>.

Fendt, M. and Fanselow, M. S. [1999]: ‘The Neuroanatomical and Neurochemical Basis of Conditioned Fear’, *Neuroscience and Biobehavioral Reviews*, **23**, pp. 743–60.

Fibiger, C. H. [2012]: ‘Psychiatry, the Pharmaceutical Industry, and the Road to Better Therapeutics’, *Schizophrenia Bulletin*, **38**, pp. 649–50.

Fried, E. I. and Cramer, A. O. J. [2017]: ‘Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology’, *Perspectives on Psychological Science*, **12**, pp. 999–1020.

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**, pp. 438–56.

Greek, R. and Shanks, N. [2011]: ‘Complex Systems, Evolution, and Animal Models’, *Studies in History and Philosophy of Biological and Biomedical Sciences*, **42**, pp. 542–44.

Gururajan, A., Reif, A., Cryan, J. F. and Slattery, D. A. [2019]: ‘The Future of Rodent Models in Depression Research’, *Nature Reviews Neuroscience*, **20**, pp. 686–701.

Hamm, A. O., Weike, A. I., Schupp, H. T., Treig, T., Dressel, A. and Kessler, C. [2003]: ‘Affective Blindsight: Intact Fear Conditioning to a Visual Cue in a Cortically Blind Patient’, *Brain*, **126**, pp. 267–75.

Hamm, A. O., Richter, J., Pané-Farré, C., Westphal, D., Wittchen, H.-U., Vossbeck-Elsebusch, A. N., Gerlach, A. L. et al. [2016]: ‘Panic Disorder with Agoraphobia from a Behavioral Neuroscience Perspective: Applying the Research Principles Formulated by the Research Domain Criteria (RDoC) Initiative’, *Psychophysiology*, **53**, pp. 312–22.

Harro, J. [2019]: ‘Animal Models of Depression: Pros and Cons’, *Cell and Tissue Research*, **377**, pp. 5–20.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. and Rosenthal, R. [2014]: ‘Clinical Development Success Rates for Investigational Drugs’, *Nature Biotechnology*, **32**, pp. 40–51.

Hawkins, J., Hicks, R. A., Phillips, N. and Moore, J. D. [1978]: ‘Swimming Rats and Human Depression’, *Nature*, **274**, p. 512.

Hollenstein, T. and Lanteigne, D. [2014]: ‘Models and Methods of Emotional Concordance’, *Biological Psychology*, **98**, pp. 1–5.

Hollis, F. and Kabbaj, M. [2014]: ‘Social Defeat as an Animal Model for Depression’, *ILAR Journal*, **55**, pp. 221–32.

Hyman S. E. [2012]: ‘Revolution Stalled’, *Science Translational Medicine*, **4**, available at <doi.org/10.1126/scitranslmed.3003142>.

Hyman, S. E. [2013]: ‘Psychiatric Drug Development: Diagnosing a Crisis’, *Cerebrum*, **2013**, available at <www.ncbi.nlm.nih.gov/pmc/articles/PMC3662213/>.

Ioannidis J. P. [2016]: ‘Why Most Clinical Research Is Not Useful’, *PLOS Medicine*, **13**, available at <doi.org/10.1371/journal.pmed.1002049>.

Kalueff, A. V., LaPorte, J. L., Murphy, D. L. and Sufka, K. [2008]: ‘Hybridizing Behavioral Models: A Possible Solution to Some Problems in Neurophenotyping Research?’, *Progress in Neuro-psychopharmacology and Biological Psychiatry*, **32**, pp. 1172–78.

Kim, C. H., Hvoslef-Eide, M., Nilsson, S. R. O., Johnson, M. R., Herbert, B. R., Robbins, T. W., Saksida, L. M., Bussey, T. J. and Mar, A. C. [2015]: ‘The Continuous Performance Test (rCPT) for Mice: A Novel Operant Touchscreen Test of Attentional Function’, *Psychopharmacology*, **232**, pp. 3947–66.

Kozak, M. J. and Cuthbert, B. N. [2016]: ‘The NIMH Research Domain Criteria Initiative: Background, Issues, and Pragmatics’, *Psychophysiology*, **53**, pp. 286–97.

Lachowicz, J., Niedziałek, K., Rostkowska, E., Szopa, A., Świąder, K., Szponar, J. and Serefko, A. [2021]: ‘Zebrafish as an Animal Model for Testing Agents with Antidepressant Potential’, *Life*, **11**, available at <doi.org/10.3390/life11080792>.

Lapate, R. C., Rokers, B., Tromp, D. P. M., Orfali, N. S., Oler, J. A., Doran, S. T., Adluru, N., Alexander, A. L. and Davidson, R. J. [2016]: ‘Awareness of Emotional Stimuli Determines the Behavioral Consequences of Amygdala Activation and Amygdala–Prefrontal Connectivity’, *Scientific Reports*, **6**, pp. 1–16.

Lang, P. J. and Davis, M. [2006]: ‘Emotion, Motivation, and the Brain: Reflex Foundations in Animal and Human Research’, *Progress in Brain Research*, **156**, pp. 3–29.

LeDoux J. [2012]: ‘Rethinking the Emotional Brain’, *Neuron*, **73**, pp. 653–76.

LeDoux, J. E. [2015]: *Anxious: Using the Brain to Understand and Treat Fear and Anxiety*, New York: Penguin Books.

LeDoux, J. E. and Pine, D. S. [2016]: ‘Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework’, *American Journal of Psychiatry*, **173**, pp. 1083–93.

LeDoux, J. E. and Brown, R. [2017]: ‘A Higher-Order Theory of Emotional Consciousness’, *Proceedings of the National Academy of Sciences USA*, **114**, available at <doi.org/10.1073/pnas.1619316114>.

López-Muñoz, F., D’Ocón, P., Romero, A., Guerra, J. A. and Álamo, C. [2022]: ‘Role of Serendipity in the Discovery of Classical Antidepressant Drugs: Applying Operational Criteria and Patterns of Discovery’, *World Journal of Psychiatry*, **12**, pp. 588–602.

Michel, M. [2019]: ‘The Mismeasure of Consciousness: A Problem of Coordination for the Perceptual Awareness Scale’, *Philosophy of Science*, **86**, pp. 1239–49.

Michel, M. [2023]: ‘Calibration in Consciousness Science’, *Erkenntnis*, **88**, pp. 829–50.

Miller, G. [2010]: ‘Is Pharma Running out of Brainy Ideas?’, *Science*, **329**, pp. 502–4.

Molendijk, M. L. and de Kloet, E. R. [2015]: ‘Immobility in the Forced Swim Test Is Adaptive and Does Not Reflect Depression’, *Psychoneuroendocrinology*, **62**, pp. 389–91.

Molendijk, M. L. and de Kloet, E. R. [2019]: 'Coping with the Forced Swim Stressor: Current State-of-the-Art', *Behavioural Brain Research*, **364**, available at <doi.org/10.1016/j.bbr.2019.02.005>.

Morris, S. E., Sanislow, C. A., Pacheco, J., Vaidyanathan, U., Gordon, J. A. and Cuthbert, B. N. [2022]: 'Revisiting the Seven Pillars of RDoC', *BMC Medicine*, **20**, available at <doi.org/10.1186/s12916-022-02414-0>.

Mul, J. D., Zheng, J. and Goodyear, L. J. [2016]: 'Validity Assessment of 5 Day Repeated Forced-Swim Stress to Model Human Depression in Young-Adult C57BL/6J and BALB/CJ Mice', *ENeuro*, **3**, available at <doi.org/10.1523/ENEURO.0201-16.2016>.

Nestler, E. J. and Hyman, S. E. [2010]: 'Animal Models of Neuropsychiatric Disorders', *Nature Neuroscience*, **13**, pp. 1161–69.

Nishimura, H., Tsuda, A., Oguchi, M., Ida, Y. and Tanaka, M. [1988]: 'Is Immobility of Rats in the Forced Swim Test "Behavioral Despair"?', *Physiology and Behavior*, **42**, pp. 93–95.

Padovani, F. [2015]: 'Measurement, Coordination, and the Relativized *a Priori*', *Studies in History and Philosophy of Modern Physics*, **52**, pp. 123–28.

Panksepp, J., Fuchs, T. and Iacobucci, P. [2011]: 'The Basic Neuroscience of Emotional Experiences in Mammals: The Case of Subcortical FEAR Circuitry and Implications for Clinical Anxiety', *Applied Animal Behaviour Science*, **129**, pp. 1–17.

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. and Schacht, A. L. [2010]: 'How to Improve RD Productivity: The Pharmaceutical Industry's Grand Challenge', *Nature Reviews Drug Discovery*, **9**, pp. 203–14.

Rachman, S. and Hodgson, R. [1974]: 'I. Synchrony and Desynchrony in Fear and Avoidance', *Behaviour Research and Therapy*, **12**, pp. 311–18.

Robinson, E. [2018]: 'Psychopharmacology: From Serendipitous Discoveries to Rationale Design, But What Next?', *Brain and Neuroscience Advances*, **2**, available at <doi.org/10.1177/2398212818812629>.

Rygula, R., Abumaria, N., Flügge, G., Fuchs, E., Rüther, E. and Havemann-Reinecke, U. [2005]: 'Anhedonia and Motivational Deficits in Rats: Impact of Chronic Social Stress', *Behavioural Brain Research*, **162**, pp. 127–34.

Schaffner, K. F. [2020a]: 'A Comparison of Two Neurobiological Models of Fear and Anxiety: A "Construct Validity" Application?', *Perspectives on Psychological Science*, **15**, pp. 1214–27.

Schaffner, K. F. [2020b]: 'Approaches to Multilevel Models of Fear: The What, Where, Why, How, and How Much?', in K. S. Kendler, J. Parnas and P. Zachar (*eds*), *Levels of Analysis in Psychopathology: Cross-disciplinary Perspectives*, Cambridge: Cambridge University Press, pp. 384–409.

Scheggi, S., De Montis, M. G. and Gambarana, C. [2018]: 'Making Sense of Rodent Models of Anhedonia', *International Journal of Neuropsychopharmacology*, **21**, pp. 1049–65.

Shanks, N., Greek, R. and Greek, J. [2009]: 'Are Animal Models Predictive for Humans?', *Philosophy, Ethics, and Humanities in Medicine*, **4**, available at <doi.org/10.1186/1747-5341-4-2>.

Shelley, C. [2010]: 'Why Test Animals to Treat Humans? On the Validity of Animal Models', *Studies in History and Philosophy of Biological and Biomedical Sciences*, **41**, pp. 292–99.

Sjoberg, E. A. [2017]: 'Logical Fallacies in Animal Model Research', *Behavioral and Brain Functions*, **13**, available at <doi.org/10.1186/s12993-017-0121-8>.

Steimer, T. [2011]: 'Animal Models of Anxiety Disorders in Rats and Mice: Some Conceptual Issues', *Dialogues in Clinical Neuroscience*, **13**, pp. 495–506.

Steru, L., Chermat, R., Thierry, B., Simon, P. [1985]: 'The Tail Suspension Test: A New Method for Screening Antidepressants in Mice', *Psychopharmacology*, **85**, pp. 367–70.

Tal, E. [2017]: 'Calibration: Modelling the Measurement Process', *Studies in History and Philosophy of Science Part A*, **65–66**, pp. 33–45.

Tal, E. [2020]: 'Measurement in Science', in E. N. Zalta (*ed.*), *The Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/fall2020/entries/measurement-science/>.

Tamietto, M., Castelli, L., Vighetti, S., Perozzo, P., Geminiani, G., Weiskrantz, L. and De Gelder, B. [2009]: 'Unseen Facial and Bodily Expressions Trigger Fast Emotional Reactions', *Proceedings of the National Academy of Sciences USA*, **106**, pp. 17661–66.

Tamietto, M. and De Gelder, B. [2010]: 'Neural Bases of the Non-conscious Perception of Emotional Signals', *Nature Reviews Neuroscience*, **11**, pp. 697–709.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M. and Lau, H. [2018]: 'Towards an Unconscious Neural Reinforcement Intervention for Common Fears', *Proceedings of the National Academy of Sciences USA*, **115**, pp. 3470–75.

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G. and LeDoux, J. E. [2022]: 'Putting the "Mental" Back in "Mental Disorders": A Perspective from Research on Fear and Anxiety', *Molecular Psychiatry*, **27**, pp. 1322–30.

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R. and Borsboom, D. [2020]: 'The Network Approach to Psychopathology: A Review of the Literature 2008–2018 and an Agenda for Future Research', *Psychological Medicine*, **50**, pp. 353–66.

Sallis, F., Lichstein, K. L. and McGlynn, F. D. [1980]: 'Anxiety Response Patterns: A Comparison of Clinical and Analogue Populations', *Journal of Behavior Therapy and Experimental Psychiatry*, **11**, pp. 179–83.

Schaffner, K. F. [2020]: 'A Comparison of Two Neurobiological Models of Fear and Anxiety: A "Construct Validity" Application?', *Perspectives on Psychological Science*, **15**, pp. 1214–27.

Seyhan, A. A. [2019]: 'Lost in Translation: The Valley of Death across Preclinical and Clinical Divide—Identification of Problems and Overcoming Obstacles', *Translational Medicine Communications*, **4**, pp. 1–19.

Tsankova, N. M., Berton, O., Renthal, W., Kumar, A., Neve, R. L. and Nestler, E. J. [2006]: 'Sustained Hippocampal Chromatin Regulation in a Mouse Model of Depression and Antidepressant Action', *Nature Neuroscience*, **9**, pp. 519–25.

Van der Staay, F. J. [2006]: 'Animal Models of Behavioral Dysfunctions: Basic Concepts and Classifications, and an Evaluation Strategy', *Brain Research Reviews*, **52**, pp. 131–59.

Van Fraassen, B. C. [2008]: *Scientific Representation: Paradoxes of Perspective*, Oxford: Oxford University Press.

Wolpe, J. [1958]: *Psychotherapy by Reciprocal Inhibition*, Stanford, CA: Stanford University Press.

Wolpe, J. [1969]: *The Practice of Behavior Therapy*, New York: Pergamon Press.

Zoellner, L. A., Rothbaum, B. O. and Feeny, N. C. [2011]: ‘PTSD Not an Anxiety Disorder? DSM Committee Proposal Turns Back the Hands of Time’, *Depression and Anxiety*, **28**, pp. 853–56.