European Journal for Philosophy of Science (2021) 11: 77
https://doi.org/10.1007/s13194-021-00395-0

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Reactivity in measuring depression

Rosa W. Runhardt[1]

Received: 3 November 2020 / Accepted: 24 June 2021 / Published online: 28 July 2021

Abstract

If a human subject knows they are being measured, this knowledge may affect their attitudes and behaviour to such an extent that it affects the measurement results as well. This broad range of effects is shared under the term 'reactivity'. Although reactivity is often seen by methodologists as a problem to overcome, in this paper I argue that some quite extreme reactive changes may be legitimate, as long as we are measuring phenomena that are not simple biological regularities. Legitimate reactivity is reactivity which does not undermine the accuracy of a measure; I show that if such reactivity were corrected for, this would unjustifiably ignore the authority of the research subject. Applying this argument to the measurement of depression, I show that under the most commonly accepted models of depression there is room for legitimate reactivity. In the first part of the paper, I provide an inventory of the different types of reactivity that exist in the literature, as well as the different types of phenomena that one could measure. In the second part, I apply my argument to the measurement of depression with the PHQ-9 survey. I argue that depending on what kind of phenomenon we consider depression to be (a disease, a social construction, a harmful dysfunction, or a practical kind), we will accept different kinds of reactivity. I show that both under the harmful dysfunction model and the practical kinds model, certain reactive changes in measuring depression are best seen as legitimate recharacterizations of the underlying phenomenon, and define what legitimate means in this context. I conclude that in both models, biological aspects constrain characterization, but the models are not so strict that only one concept is acceptable, leaving room for reactivity.

Keywords Reactivity · Measurement · Philosophy of social science · Philosophy of psychology · Psychometrics · Human kinds · Looping

This article belongs to the Topical Collection: Reactivity in the Human Sciences
Guest Editors: Marion Godman, Caterina Marchionni, Julie Zahle

✉ Rosa W. Runhardt
r.w.runhardt@phil.leidenuniv.nl

[1] Institute for Philosophy, Leiden University, Leiden, the Netherlands

1 Introduction

If a human subject knows they are being measured, this knowledge may affect their attitudes and behaviour to such an extent that it affects the measurement results as well. This broad range of effects is shared under the term 'reactivity'. For example, if a survey makes a breast cancer patient reflect on the less fortunate in its first few questions, the subject may rate her quality of life higher in the survey's subsequent questions (cf. Wood, 1985). Or consider a case when a subject starts walking more after a baseline interview with a researcher because they wish to impress the researcher with their newfound commitment to fitness. While frustrating for the researcher, in this paper I argue that some quite extreme reactive changes may be legitimate, as long as we are measuring phenomena that are not simple biological regularities. Applying this argument to the measurement of depression, I show that under the most commonly accepted models of depression there is room for reactivity.

Measurement involves several stages (Cartwright & Runhardt, 2014). To measure, one first needs to characterize a phenomenon, then choose an appropriate representation for the concept and finally design appropriate on-the-ground procedures. While I will show that reactivity can affect all three stages of measurement, many methodologists fixate on the procedures stage. Such methodologists often only want to fix or avoid reactivity, thinking it biases their research findings (French & Sutton, 2011, 273). They emphasize designing new types of procedures; for example, then-tests, structural equations modelling, and latent trajectory analysis to detect reactivity in quality of life research (Ahmed & Ring, 2008; Sajobi et al., 2018).

Although reactivity is seen as a problem to overcome, in this paper I argue that under certain conditions reactivity is a justifiable change in the characterization or representation steps of measurement. While such cases of reactivity change the measure (potentially making it incomparable to previous results) this change is nevertheless *legitimate*. As I will define and defend in more detail below, by 'legitimate' I mean specifically that it does not undermine the accuracy of the measure. Respecting such reactivity grants a research subject some authority in the measurement process. Whether reactivity is legitimate in this narrow sense, I will argue, depends on the type of phenomenon we are measuring. To organize this argument, I provide an inventory of both the different types of reactivity and the different types of phenomena that one could measure.

Depression is an exemplary case for analysing reactivity in measurement; it is inextricably linked to a person's mood state, which we cannot measure without the person being aware of the measurement (French & Sutton, 2010). Researchers often have to rely on self-reports for their measurement of such disorders, but self-reports are susceptible to reactivity: "measurements can actively interfere with the observed phenomenon, thus not only imperfectly capturing the psychological experience but potentially contaminating it indelibly" (Johar & Sackett, 2018, 304).

I will analyse reactivity in the measurement of depression with the PHQ-9 survey, which is based on the DSM-IV classification of depressive disorders. I argue that depending on what kind of thing we consider this mental disorder to be (a disease, a social construction, a harmful dysfunction, or a practical kind), we will accept different kinds of reactivity. I show that both under the harmful dysfunction model and the

practical kinds model, certain reactive changes in measuring depression are best seen as legitimate recharacterizations of the underlying phenomenon. I conclude that in both models, biological aspects constrain characterization, but the models are not so strict that only one way of characterizing the phenomenon is legitimate.

The legitimacy of reactivity in measuring depression requires that reactivity should not be corrected for by researchers, as this would ignore the authority of the research subject. In particular, the research subject should be given a certain amount of authority to control their characterization and representation of depressive phenomena, as long as the characterization and representation are compatible with the aforementioned biological constraints. If compatible, reactivity does not undermine the accuracy of the measure. Calling reactivity 'legitimate' does not mean, here, that researchers do not need to find out whether reactivity occurs. Rather, finding out more about the changing characterization and representation of depression by individuals is a productive area of research.

My analysis clarifies the parallels between the study of reactivity in the methodological literature on the one hand, and Jonathan Tsou's recent philosophical interpretations of Ian Hacking's theories of human kinds on the other (Hacking, 1995, 1999; Tsou, 2007, 2016, 2019). Ian Hacking's original concept of 'looping', i.e. the idea that classifying people changes them to such an extent that researchers have to rethink their classifications, (Hacking, 1995, 369), is a type of reactivity as I have defined it above. Like Tsou and Hacking, methodologists studying psychiatric classification increasingly highlight the interplay between biological and social factors. Tsou focuses in particular on "objects of classification (i.e., kinds of people) that can be identified with reference to a law-like biological regularity and are aware of how they are classified" (Tsou, 2007, 329). Tsou's discussion of this 'awareness' is still limited. By discussing different types of reactivity, this paper expands on Hacking's and Tsou's analyses, bridging the gap between the methodological literature and philosophical work on human kinds. I will briefly discuss the parallels between my argument and this literature near the end of the paper.

The paper, then, is set up as follows. In the first part of the paper, I present a general theory of measurement and reactivity that forms the foundation for my subsequent arguments. In the second part of the paper, I introduce the PHQ-9 questionnaire for measuring depression, as well as the four main theories of what kind of phenomenon depression is, viz. the disease model, social constructionist model, harmful dysfunction model, and practical kinds model. I show that each of these four models handles reactivity differently. I give an inventory of which occurrences of reactivity are legitimate under each model. Finally, I conclude that under the most promising of the four theories, extensive reactive change is legitimate, and briefly compare this conclusion to the Hacking/Tsou analysis of psychiatric classification.

2 A general theory of measurement

Cartwright & Runhardt (2014)[1] describe three phases for measurement in the social sciences: characterization, representation, and procedures. Characterization here means delineating the phenomenon concerned in such a way that it is fit

[1] See also Cartwright, Bradburn, and Fuller(2017).

for measurement. Representation means choosing a scale (nominal, ordinal, interval, ratio) or table of indicators with which to represent the measurement outcome. Finally, procedures in measurement are those on-the-ground methods we use to come to conclusive figures, such as a ranking on the chosen scale or a filled in table.

While for some concepts, only one way of characterizing will be correct (e.g. classic 'natural kinds' like gold), phenomena best described with what Cartwright and Runhardt call 'Ballung concepts' are different.[2] Cartwright and Runhardt say such phenomena "are characterised by family resemblance between individuals rather than by a definite property" (Cartwright & Runhardt, 2014, 268),[3] and argue Ballung phenomena are rife in the social sciences. For Ballung phenomena, characterization necessarily involves choices in the definition of the concept; the concept must be more strictly circumscribed than the phenomenon to make measurement possible, but there is no single right way of doing so. Because of this gap between concepts and phenomena, a social scientist will often have some 'wiggle room' in choosing the appropriate characterization, representation, and procedures for measurement. As a result, often several different acceptable measures exist for one Ballung phenomenon. As long as the three stages of measurement are internally consistent, Cartwright and Runhardt argue, there is no reason to prefer one measure over another.[4]

Cartwright and Runhardt make a clear case for why e.g. phenomena in political science, like 'civil war', are Ballung phenomena. However, it is too quick to assume all human phenomena[5] will fit the authors' framework. For example, if we wish to extend Cartwright and Runhardt's argument to mental disorders, we must first take a position in the debate about what kinds of phenomena mental disorders are; it is not at all clear that all mental disorders are Ballung phenomena. More subtle distinctions will be necessary which the idea of a Ballung phenomenon does not capture.

[2] In the article, the authors do not make an explicit distinction between phenomena ('out there' in the social world) and concepts (human constructions). However, I will do so in this paper.

[3] The term 'Ballung' is derived from the work of Otto Neurath; see Cartwright and Bradburn 2010 for a longer analysis of the term. An abundance of similar terms exist, from cluster concepts (Daniel Little; these concepts correspond to "a variety of phenomena that share some among a cluster of properties" (Little, 1993, 190)) to human kinds (Ian Hacking; for these kinds, the meaning of the concept influences the individuals falling under it, and vice-versa, in the 'looping effect' mentioned in the introduction of this paper), to nomadic concepts (Catherine Greene; these concepts are cluster concepts with the added difficulty that the boundaries which phenomena the nomadic concept represents "change over time" (Greene, 2019, 11)).

[4] The authors stress that these three phases need to 'mesh', i.e. be mutually consistent. Some characterizations will lend themselves more to a certain type of representation, for instance. If, in the characterization stage, one decides one is dealing with a multifaceted concept, a table of indicators may be more appropriate as representation. If we find that certain procedures are impossible, we might go back to the drawing board and settle on a different characterization. Moving beyond Cartwright and Runhardt's original analysis, we may argue for additional requirements on measures. For example, one might argue that an internally consistent measure is only of use if it has some relevant joint explanatory or predictive power for the purposes for which the measure is designed.

[5] In calling these phenomena 'human', I have in mind something akin to Hacking's original definition of human kinds (cf. Hacking, 1995). I wish to sidestep any debates here about whether such phenomena are only studied in the social sciences and whether e.g. psychometrics is exclusively a part of social science. I do, however, wish to limit this paper to the discussion of reactivity, which under my definition only occurs when measuring human subjects; as such, focusing on human phenomena is appropriate.

Fortunately, Cartwright and Runhardt's framework is useful beyond Ballung phenomena. There are other types of phenomena that can be characterized in multiple ways: chief among them those phenomena that researchers approach by developing *thick* concepts, i.e. concepts which combine evaluative and descriptive elements.[6] The idea of thick concepts originally came from moral philosophy (Williams, 1985) and is now applied to understanding measurement in philosophy of social science by amongst others Anna Alexandrova (Alexandrova, 2018). A good example is Alexandrova's study of well-being, where she argues that "'well-being' is thick to the extent that it is a good thing to have, but also to fare well is to have a certain amount of health, not to be depressed, lonely, and so on" (Alexandrova, 2018, 425). Alexandrova calls hypotheses about thick concepts 'mixed', defining them as follows:

> "A hypothesis is mixed if and only if:
>
> 1 It is an empirical hypothesis about a putative causal or statistical relation.
> 2 At least one of the variables in this hypothesis is defined in a way that presupposes a moral, prudential, political, or aesthetic value judgement about the nature of this variable." (Alexandrova, 2018, 424)

Since thick concepts contain evaluative elements, their characterization will involve evaluative choices in the definition of the concept, just like was the case for Ballung concepts. Such characterizations will then have to mesh with an appropriate representation and characterization. In short, several different acceptable measures will exist for thick concepts.

In Section 4, I will show that the role reactivity can play in measurement in psychology depends on the status of the phenomenon involved: are we dealing with a disease, a social construction, a harmful dysfunction, or a practical kind? The theory above will allow us to better understand these different options. First, however, I will use Cartwright and Runhardt's framework to distinguish the different types of reactivity one can encounter when measuring human phenomena.

3 A trichotomy of reactive change

As stated in the introduction, I define reactivity broadly, as those cases in which a human subject knows they are being measured and this awareness affects their attitudes and behaviour to such an extent that it affects the measurement results.

[6] Note that the set of phenomena we pick out with thick concepts are not identical to the set of Ballung phenomena. What is required for a concept to be thick is that it includes some evaluative judgement; Ballung phenomena, however, do not necessarily require a researcher to make evaluative judgements. Moreover, while Ballung phenomena are characterized by family resemblance rather than one definitive property, there is no such requirement on phenomena picked out by thick concepts.

However, different subtypes of reactivity in measuring human subjects exist. Following Robert Golembiewski and colleagues (Golembiewski et al., 1976), I will call these subtypes alpha, beta, and gamma change.[7]

3.1 Alpha change

Alpha change is the simplest type of reactive change in the Golembiewski trichotomy, namely a straightforward change in the numerical outcome of a measurement due to being measured. Think back to one of the examples of reactivity in the introduction: a subject increases their daily step count to impress a researcher with their fitness. We can verify this change in step count, e.g. with a smartphone (cf. Glynn et al., 2014). Imagine that before meeting a researcher for the first time, a subject's step count was collected automatically; the subject is measured to take an average of 3,000 daily steps. The researcher and subject meet have a baseline interview, in which this step count is discussed. After this interview, the step count continues to be collected and during a follow-up interview a few weeks later, the same person totals an average of 4,900 daily steps. As long as the measurement tool measures in the same way during both measurement periods (e.g. in both periods, the subject kept the smartphone on their person all day), this higher average step count is a case of alpha change.[8] While such a higher average can prove a knotty research problem, the alpha reactive effect is philosophically unproblematic, since it is ultimately verifiable that the different step count was the result of being measured.[9]

3.2 Beta change

Beta change, in contrast, is more complex. To illustrate, consider the other example of reactivity from the introduction: a breast cancer patient reports a higher quality of life if the survey she is filling out first prompts her to make 'downward comparisons' to patients worse off than her (Wood et al., 1985). While complicating the researcher's life considerably, I would argue such a move is as *legitimate*

[7] Methodologists (Norman and Parker, 1996; cf. French and Sutton, 2010) amongst others use the Golembiewski trichotomy to assess reactivity in quality of life research (Ahmed and Ring, 2008; Schwartz and Sprangers, 1999; Sprangers and Schwartz, 1999). An alternative term in this literature is 'response shift', "changes in the meaning of one's self-evaluation of a target construct" (Sprangers and Schwartz, 1999, 1508), particularly after life-altering illness.

[8] We can also measure fitness (less exactly) with a self-report; if the self-reported amount of walking has gone up from the first to the second time of reporting, this is alpha change. Note that we should distinguish between cases of alpha change where a subject is intentionally misreporting their step count versus cases where they are reactive, but honest. Intentional misreporting (e.g. a subject says they meet the 10,000 daily step count to get the researcher off their back, but in fact they lead a sedentary lifestyle) will not form part of this paper, because such reactive changes are nearly always illegitimate for purposes of research.

[9] Not all alpha change as defined by Golembiewski et al. is reactive change; a person's average daily step count may change because of influences unrelated to being measured, e.g. a change from a mostly sedentary job to a more active one.

as the alpha change above. Recall that 'legitimate' here means that the reactive change does not undermine the accuracy of the measure.[10] In respecting reactivity, the patient is given a certain amount of authority: she is allowed to change what she considers her quality of life to be, e.g. by changing how she grades her own quality of life; I would argue we cannot accuse her of 'getting it wrong'. Empirical evidence indeed supports that, like in our example, "selectively comparing downward (…) [can contribute] to a feeling of relative well-being" (VanderZee et al., 1995, 453). In general, if social comparisons are part of our overall measurement tool (e.g. part of the survey we give), then such comparisons may impact on the quality of life a person experiences, and thus reports (VanderZee et al., 1995; cf. Gibbons, 1999; Sprangers & Schwartz, 1999).

In the cases described above, a test subject will report a shift because they interpret the scale in a different way; the subject 'recalibrates' the scale they have been using. As such, I consider beta change to be a change in Cartwright and Runhardt's representation stage of measurement: in this stage, one 'marks' the scale chosen with some standard measurement results.[11] If a subject has so far felt her quality of life ranks at the bottom of the scale, a survey which lets her find out 'things could be worse' will lead to a recalibration. As Golembiewski et al. argue, a subject's experiences will act as "anchoring points" (Golembiewski et al., 1976, 136) for their (re)calibration; taking part in a measurement could be such an 'experience', and so recalibration may be a case of reactivity.[12]

Importantly, the reason beta change is legitimate in the measurement of quality of life is because of the phenomenon characterized as 'quality of life' requires a thick concept: it contains both descriptive and evaluative elements. In the terms of the previous section, the respondent's recalibration of quality of life is acceptable because of the concept's 'thick' status, as long as it is internally consistent. Contrast this example with the daily step count example from our introduction; one's step count is not a thick phenomenon, and it is possible to calibrate this measurement

[10] Note that my concept of legitimate reactivity differs considerably from María Jiménez-Buedo's concept of benign reactivity (Jiménez-Buedo, 2021). For Jiménez-Buedo, "[b]enign reactivity occurs when the intervention's impact on the subject's behavior does not affect the output variable of interest in the experiment" (Jiménez-Buedo, 2021, 13). She contrasts this with malignant reactivity, which occurs "when the experimental manipulation not only changes the value of the putative effect Y by setting in motion the putative cause X, but additionally, it gives rise to an additional causal path that also affects the output variable of interest Y" (Jiménez-Buedo, 2021, 13). However, in deciding whether beta change is legitimate, we are concerned with a potential change in how we assign a number to the variable Y in the first place, and not with the potential addition of causal paths. We may say that for legitimacy, it is only the *reference* of variable Y that is at stake. In the Woodwardian framework, we ought to delineate X and Y (and thereby, establish what they refer to) carefully before establishing the effect of a putative intervention I on the $X \rightarrow Y$ relationship (cf. Woodward, 2003, 115). In sum, I would argue that legitimate reactivity and benign reactivity are qualities that refer to different stages of the research process.

[11] Cartwright and Runhardt also mention that the choice of representation could be not for a scale, but rather for a table of indicators. I will ignore this aspect of representation here.

[12] Not all beta change will be reactive. Other experiences a subject may have between the first and second time they are being measured can lead to a recalibrated scale.

in the wrong way (e.g. by using a monitor that is too sensitive).[13] In that case, beta change would be *illegitimate,* since it does undermine the accuracy of the measure. Speaking bluntly, the subject should not get the authority to recalibrate the monitor.

3.3 Gamma change

Gamma change, finally, is a change in characterization. For example, if the breast cancer patient in our earlier example were to construct a different meaning of the concept 'quality of life' between measurements, and this would lead to a different result in the second measurement, Golembiewski et al. would call this gamma change. "Gamma change involves a redefinition or recharacterization of some domain, a major change in the perspective or frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality." (Golembiewski et al., 1976, 135) When this recharacterization is a result of being measured, the gamma change is a case of reactivity.

I interpret gamma change as a change in the characterization stage of measurement. As such, gamma change can have serious results for the comparability of the results of measurement before and after the change has taken place. If the research subject self-reports a certain value of quality of life the first time she is measured, then changes her mind about what quality of life means to them and as a result reports a different value of quality of life the second time, the two results are incomparable. In later sections, I will discuss when gamma change is legitimate, and what its effects will be on researchers' inferences.

4 Reactivity in measuring depression

We have so far seen that we can divide reactive changes in measurement results in three categories: alpha change, beta change, and gamma change,[14] where beta change impacts on the representation stage of measurement and that gamma change impacts on the characterization stage. In this section, I will apply this trichotomy to the literature on depression.

Several empirical studies show that reactivity affects the measurement of negative mood states such as depression. Quantitative studies (cf. Choquette and Hesselbrock (1987) and Sharpe and Gilbert (1998)[15]) report that repeatedly administering some measures of negative mood states has a (statistically significant) reactive effect on reported depression, controlling for all other influences. Methodologists warn of the potential damage such reactivity might have: "differences [caused by reactive bias] could easily mislead a researcher into believing [depression] was alleviated when in fact it was not" (Choquette & Hesselbrock, 1987, 277), with serious issues for further research and treatment.

[13] It is also not a Ballung phenomenon (not characterized by family resemblance).

[14] In more complex situations, several types of change may influence the same measurement (cf. Sprangers and Schwartz, 1999). I will not discuss this complication in the paper; following Cartwright and Runhardt, as long as the new measure 'meshes' this will not affect my argument.

[15] See French and Sutton (2010) for a further overview.

As stated earlier, however, I will argue that not all reactivity in measuring depression is unacceptable; some reactive change may be a legitimate recalibration or recharacterization by the research subject. To make this argument, let me outline a concrete research paper on depression as an example. I will show that this research paper contains the different types of reactivity outlined above.

Carina Marsay and her colleagues (Marsay et al., 2018) report a qualitative methods study of antenatal anxiety and depression in South-Africa, where poor urban women attending a high-risk antenatal clinic participated in a screening interview for mental illness during the second trimester of their pregnancy. When the same women were re-interviewed in the third trimester, they often reported lower levels of antenatal depression. Marsay et al. report that many of the women partially attribute this drop in levels to "changes in emotion, cognition and behaviour that occurred as a result of participating in the screening interview" (Marsay et al., 2018, 352); in other words, the change in levels was a reactive effect. Some women say that the simple act of telling the interviewer how they were feeling had a therapeutic effect. Others said that the interview made them more aware of their mental health, which led them to seek out social support or otherwise make changes that helped them cope better. In sum, being interviewed and measured led to a range of effects, including "gaining self-knowledge, validation of experiences and personal agency, and seeking out support from others" (Marsay et al., 2018, 357).

We can categorize the reactive changes Marsay et al. postulate using the reactive change trichotomy from the previous section. For example, if the interview itself led a woman to seek social support, her depression may become less severe. This change would be alpha change: it is not due to a change in how the woman characterizes depression, nor due to a recalibration of her scale of depression. On the other hand, if a woman gains self-knowledge of her mental health during and after the first interview, and subsequently reports a different level of depression the second time she is interviewed, this change may well be beta or gamma change. She may be better able to compare her situation to the situation of others (recalibration) or better able to put her own feelings into words (recharacterization).

4.1 The PHQ-9

Now that I have illustrated my terminology with Marsay et al.'s study of antenatal depression, I will turn to a more extensive example. In the remainder of this paper, I will focus on the PHQ-9, a patient questionnaire for measuring the severity of depression, first used in primary care settings. The PHQ-9, developed by Robert Spitzer, Janet Williams, and Kurt Kroenke (Kroenke & Spitzer, 2002), is part of the broader Patient Health Questionnaire (PHQ), which in turn is a self-report version of the broader Primary Care Evaluation of Mental Disorders (PRIME-MD) diagnostic tool. We can distinguish the three stages of measurement from Cartwright and Runhardt (2014), i.e. characterization, representation, and procedures, in the PHQ-9.

4.1.1 Characterization

The PHQ-9 *characterizes* depressive phenomenon based on the DSM-IV, the fourth edition of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (APA, 2000). The DSM-IV gives 9 symptoms for depression, which each get their own question in the PHQ-9, asking the respondent to self-report on that aspect of their lives in the last two weeks:

> "(1) Little interest or pleasure in doing things; (2) Feeling down, depressed, or hopeless; (3) Trouble falling or staying asleep, or sleeping too much (4) Feeling tired or having little energy; (5) Poor appetite or overeating; (6) Feeling bad about yourself – or that you are a failure or have let yourself or your family down; (7) Trouble concentrating on thing, such as reading the newspaper or watching television; (8) Moving or speaking so slowly that other people could have noticed. Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual; (9) Thoughts that you would be better off dead or hurting yourself in some way" (Kroenke & Spitzer, 2002, sec. Appendix).

The Diagnostic and Statistical Manual has been based on such 'diagnostic criteria' since its third edition, the DSM-III (APA, 1980).[16]

4.1.2 Representation

A subject answers each of the 9 questions in the PHQ-9 by ranking how often they were bothered by the symptom for the last two weeks on a scale of 0 (not at all) to 3 (nearly every day). The clinician or researcher then adds all 9 figures to calculate the measurement result. As such, the PHQ-9 *represents* the concept depression on a ratio scale of 0 to 27: one could order individual responses to the questionnaire on this scale (a higher number on the scale means the individual is more depressed), and the scale also has a natural zero point. The numerical score is subsequently used to categorize the individual on a more basic ordinal scale[17]: a score of 0–4 is labelled none-minimal, 5–9 is labelled mild, 10–14 is labelled moderate, 15–19 is labelled moderately severe, and 20–27 is labelled severe.[18]

[16] See also Horwitz (2002), Horwitz & Wakefield (2007), and Tsou (2016; 2019) for brief overviews of the history of the DSM and the non-scientific influences on its development. The DSM-5 (first with an Arabic numeral in its title) has since been published (APA, 2013). In this paper, I will not discuss the DSM-5, since I focus on the PHQ-9, which was based on the DSM-IV. The symptoms for depression in the DSM-IV and DSM-5 are the same. The only change is that the recently bereaved are no longer excluded from being classified as depressed in the DSM-5; this change, however, is beyond the scope of this paper, as it is irrelevant to reactivity concerns.

[17] One of the issues in representation is how many categories we ought to distinguish in measuring a mental disorder like depression. The other is how to calibrate the scale. I will focus on the latter in this paper, since reactive beta change is typically a recalibration.

[18] The diagnostic criteria for each disorder in the DSM sometimes act as a set of necessary and sufficient conditions for being diagnosed with that disorder, but the case of depression is more subtle. If a patient displays 5 or more of the 9 symptoms for 2 or more weeks, including depressed mood or anhedonia, and the symptoms cause clinically significant distress or impairment, this implies a diagnosis of Major Depressive Disorder. See Smarr (2003).

4.1.3 On the ground procedures

On the ground measurement procedures differ widely. The PHQ-9 was developed for primary care physicians, who can ask a patient to fill out the questions in self-administered written form (e.g. with paper/pencil or computer) or interview them (e.g. by telephone or in person). The PHQ-9 is now also used beyond primary care settings, in more specialized medical centres like oncology and gynaecology clinics (Smarr, 2003). Its original English version has been translated to over 30 other languages (APA, 2011).

4.2 Four models of the PHQ-9's analysis of depression

As stated above, the characterization stage of the measurement process is the delineation of a phenomenon in such a way that it is fit for measurement. In this stage the PHQ-9 relies on the DSM-IV's delineation of mental phenomena into the concept depression, as it uses the exact 9 symptoms stated in the DSM-IV. The literature contains four main positions on the relation between the symptoms in the DSM-IV and the underlying depressive phenomena: (1) the disease model; (2) the social constructionist model; (3) the harmful dysfunction model; and finally (4) the practical kinds model and its recent 'upgrade', the mechanistic property cluster model.

Each model will handle the three kinds of reactivity from Section 3 differently. As I will show below, the disease model is an essentialist natural kinds model, based on the assumption that only one characterization of depressive phenomena is correct and as such will reject beta and gamma reactive change as illegitimate, since these changes undermine the accuracy of the measure. Under the social constructionist model, on the other hand, a variety of ways of characterizing depressive phenomena are acceptable, constrained only by what characterizations are accepted in the community. Both positions have received extensive criticism; two more recent alternatives, the harmful dysfunction model and the practical kinds model, show that we can characterize the phenomena in several acceptable ways, constrained by biological aspects of the phenomena. I will show that under those more recent positions, all three types of reactive change may be legitimate in the narrow sense introduced above. Under those positions, the research subject should be given a certain amount of authority to control their characterization and representation, as long as they are compatible with the biological restraints on the phenomenon. I will now begin by outlining each of the four models, before positing how each model deals with reactivity in judging the appropriateness of the PHQ-9.

4.2.1 The disease model of mental illness

The PHQ-9 and DSM-IV characterization of depression is said to rely on the so-called 'disease model' or 'natural kinds model' of mental illness (Horwitz, 2002; Horwitz & Wakefield, 2007; Horwitz, 2014; Zachar, 2014), according to which a simple underlying phenomenon in nature which produces the 9 symptoms listed,

described as an "objective natural entit[y]" (Horwitz, 2002, 4.5) or "underlying essence" (Kendler et al., 2011, 1144) which has "fixed internal properties" (Zachar, 2000, 168). This suggests that there is only one correct way to delineate depression for psychometrics, namely by tracking this natural entity. If we are able to find this correct delineation, our measurements will be perfectly reliable.

Since the disease model assumes that a mental disorder concept like depression must track an 'objective natural entity', this model does not leave room for a reactive recharacterization or recalibration of the concept depression. To see this is the case, assume that the characterization used in some measurement *M* (say, the 9 symptoms in the PHQ-9) is the best way to track this essence; *M* is a valid and reliable measure of the intended disease model conception of depression. In that case, a beta or gamma change in the second measurement would be illegitimate, since it would mean the measure is no longer valid and reliable. If one were to recharacterize (aspects of) depression, one would no longer accurately track the essence. As such, the only acceptable reactivity in the disease model of depression is alpha change.[19] Any other type of reactivity leads to false measurement results. The disease model of depression is not commonly accepted in the current philosophical literature (Zachar, 2000; Horwitz 2002; Cooper, 2004; Horwitz and Wakefield, 2007; Kendler et al., 2011; Tsou, 2016). Amongst others, critics argue that the phenomena characterized as 'depression' by the DSM-IV are 'fuzzier' than the disease model accepts. The individuals that the DSM-IV would label as depressed vary. Lumping these individuals under the same label would not respect this variety.[20] Moreover, a person can be depressed due to different (potentially probabilistic) causal factors, in different contexts, only some of which the DSM-IV captures (cf. Kendler et al., 2011). As such, critics argue that the DSM's category for depressive disorders does not correspond to a single 'essential', 'objective' natural phenomenon (cf. Tsou, 2019). Let us now turn to some of the proposed alternatives to the disease model.

4.2.2 The social constructionist model

At the opposite side of the spectrum from the disease model is the social constructionist view. For the social constructionist, depression is not a natural entity

[19] Consider, now, the alternative, a measurement $${M}^{*}$$ which is not valid and reliable for the intended disease model conception of depression. It may be that through recharacterizing aspects of $${M}^{*}$$, the measurement properties of $${M}^{*}$$ improve because it will more accurately track the essence than before. In other words, if the characterization of $${M}^{*}$$ is not yet tracking the natural entity accurately, beta or gamma shift could lead to an improvement in measurement. However, such change is not typically what we are concerned with when we discuss reactivity.

[20] In particular, Horwitz and Wakefield argue that the DSM-IV cannot adequately distinguish between people who have (a subset of) the 9 symptoms without while their brain function is normal (e.g. people who have some symptoms because they have experienced a great loss, which Horwitz and Wakefield call a case of 'normal sadness') and those people who are symptomatic due to a dysfunction. The distinction is relevant because while the symptoms of depression and normal sadness are the same, Horwitz and Wakefield claim that their causes generally differ. Thus, making the distinction is crucial for causal analysis in depression research. See the discussion of Horwitz and Wakefield's harmful dysfunction model below.

at all: instead, it is constructed by "social systems of meaning" (Horwitz 2002: 5.6). Researchers measure and analyse psychiatric kinds like depression because these kinds are accepted by and of interest to the society these researchers are a part of (cf. Kendler et al., 2011). However, the DSM's characterization is as valid as other cultural constructions, from "unconscious forces" to "demonic possession" (Horwitz, 2002: 8.9). Different ways of characterizing depressive phenomena are acceptable, including the DSM-IV symptoms; all that is required of a characterization of depression is that it is sufficiently wide-spread or stable to be regarded as accepted by a community.

The social constructionist view, too, has direct implications for how one ought to treat reactivity in measurement. Reactivity could be just one more change due to a shift in social circumstances. So, for example, if one changes one's mind about the meaning of anhedonia (i.e. little interest or pleasure in doing things) because of participating in the PHQ-9, then that is acceptable. In this view of depression, there is no underlying entity being traced. A person will give answers to the survey that are true for the community they find themselves in. Measurement results might change because of alpha change (e.g. because the subject has sought further social support after recognizing their initial high score on the PHQ-9 and this support made them less depressed). On the other hand, measurement results might also change because the subject (or their community) has changed their interpretation of what is being asked in the questionnaire. While a careful study of this interplay between subject, community, and concept is beyond the scope of this paper, we can draw the preliminary conclusion that beta change and gamma change would not be illegitimate for the social constructionist, since there is no such thing as 'truth to nature' of the measurement.

But while the disease model of depression is too strict in its insistence that the 9 symptoms of the DSM pick out one exact 'objective, natural entity' or 'essence', the social constructionist view is often thought of as too lenient. For one, this model ignores the empirical evidence that shows the PHQ-9 functions the same across different cultures (cf. Huang et al., 2006; Kendler et al., 2011). Let me now present two more recent alternative models, which attempt to balance the disease model's search for an 'objective natural entity' on the one hand, with the social constructionist's highlighting of social influences on depression on the other.

4.2.3 The harmful dysfunction view

One answer to the tension between biological and social aspects of mental disorders is the 'harmful dysfunction view', due to Jerome Wakefield (1992) and explored in detail for depression together with Allan Horwitz (Horwitz and Wakefield 2007). According to this model, a mental disorder is a dysfunction (a breakdown of some natural system in the brain) that harms the person's well-being. Wellbeing here is "defined by social values and meanings" (Horwitz & Wakefield, 2007, 17). Therefore, whether we consider a dysfunction to be harmful, and thereby a case of depression, depends on these 'social values and meanings' as well. Since mental disorders contain both descriptive and evaluative

components in the harmful dysfunction view, we can conclude that for this view, mental disorders are a thick concept, as defined in Sect. 2.

Some aspects of the PHQ-9 measure whether the symptoms a person has are harmful or not. For example, the PHQ-9 asks explicitly whether someone has been *bothered* by the symptoms in the last 2 weeks, not just whether someone has the symptoms. I would argue that the social values discussed above may play a role in a person's 'feeling bothered'; if for whatever reason the person changes which social values and meanings they rely on in answering the question due to being measured, then this is a case of (beta or gamma) reactivity but not one that should be rejected within the harmful dysfunction model. Such reactivity is legitimate, since it does not undermine the accuracy of the measure, and so we need to respect the authority of the research subject in this change of social values and meanings.

To illustrate, imagine one of the women enrolled in Marsay et al.'s study of antenatal depression. In the first interview, she is not as aware of her mental health as she is in the second interview. She thinks she has not been bothered by the symptoms the first time she is asked. But after the interview, this woman becomes more aware of her mental state; because of this, she becomes aware of her anhedonia, her sleeplessness, etc.: the next time she is asked, she scores herself higher, as she has been bothered more. Because she was measured, she is now more aware of the harm her mental state causes. Importantly, that the reactivity is legitimate here does not mean that researchers do not need to find out whether reactivity has occurred. Rather, finding out more about the changing characterization and representation of depression by individuals is a productive area of research, and a valuable part of the Marsay et al. study.

To continue, let us reflect in more detail on the interplay between biological constraints and measurement here. The example above raises the question: can someone be depressed but unaware of it? We must distinguish here between the somatic (physical) and mental aspects of the DSM's characterization of depression. I will now make the case that reactivity affects the measurement of somatic and mental properties differently.

Arguably, we are mostly aware of somatic symptoms, such as trouble falling asleep. In fact, such physical symptoms could be seen as referring to 'objective entities': there is a fact of the matter as to how long it takes us to fall asleep. An actigraphy monitor wristwatch might log this time using a sensitive motion sensor (cf. Lauderdale et al., 2008). As such, only alpha change in this symptom would be acceptable. Other reactive changes are illegitimate.

The mental symptoms of the DSM's characterization are different. An argument from the literature on the measurement of happiness by Daniel Haybron is relevant here (Haybron, 2007). Haybron argues that humans suffer from 'affective ignorance': we are not always fully aware of how (un)happy we are or how much pleasure an activity really gives us, being unreliable in assessing affective states in general. Therefore, I would argue, self-reports about the mental symptoms of depression like the PHQ-9 are susceptible to both beta and gamma change,[21] and this change is legitimate there.

[21] See Haybron (2007, 411–413) for a discussion of the impact of affective ignorance on empirical studies of happiness.

Returning to the example, where a woman in Marsay's study becomes aware of her anhedonia because she was measured. This is not mere alpha change. Loosely speaking, this woman may have been as depressed in the first measurement as she is in the second measurement. But only now has she got the right concepts in mind to consider herself with: she becomes aware of her depression, which includes gaining the right vocabulary to talk about how she feels. As such, this is a case of gamma change. Arguably, while it complicates the researcher's inferences considerably, this change in the measurement would be legitimate under the harmful dysfunction model. Both during the first and the second measurement, the woman accurately self-reports. Both ways of considering herself are acceptable. Finding out more about the changing characterization and representation of depression by such individuals should be part of the researcher's agenda.

While some reactivity, as described in the previous example, is acceptable, the harmful dysfunction model is not as lenient as the social constructionist model. The goal of the PHQ-9 should be to find those people for whom the response is due to a dysfunction and not e.g. a bereavement or other "normal loss response" (Horwitz and Wakefield, 2007, 16).[22] Take the symptom of 'feeling down, depressed, or hopeless'. Reactivity may involve a recharacterization of what e.g. 'feeling down' means to the subject. However, in the harmful dysfunction view, the idea of feeling down is biologically stable (other animals can feel down, too) and culturally stable (antidepressants work in different cultures). A recharacterization is only valid if it respects this biologically, culturally stable 'thing' (e.g. serotonin levels). As such, the harmful dysfunction model of depression and the disease model discussed above are similar. Horwitz and Wakefield still assume there is a scientific kind 'out there in nature' to track. We may only call a person depressed if doing so respects the scientific kind, which means that not *all* reactive changes will be legitimate.

The harmful dysfunction is not the only model of mental illness meant to replace the more problematic disease model and social constructionist model. I will now turn to a final model in the literature, the practical kinds model, and its upgrade, the mechanistic property cluster (MPC) kind model. We will see that there are clear parallels between this model and the harmful dysfunction model because both models respect the interplay between biological and social properties in characterizing depression.

4.2.4 The practical kinds and mechanistic property cluster kinds models

Peter Zachar bases his practical kinds model on the notion that more than biological factors are important in characterizing a mental disorder like depression. Zachar calls his practical kinds model a pragmatist theory: he argues that different ways of delineating disorders may be valuable, depending on pragmatic considerations such as what the effects of being labelled with a disorder may be for an individual and which treatments are currently available. As such, the *utility* (broadly construed) of a classification is central to the characterization stage of measurement. In sum,

[22] See footnote 17.

according to Zachar's model different characterizations of depression will be possible, depending in part on these utility considerations.

In my discussion of Marsay et al., I discussed the example of a pregnant woman changing her characterization of anhedonia in part because of taking part in the first interview. In the example, because the woman takes part in the first interview, she has a lower score on anhedonia when she is measured with the PHQ-9 during the second interview. This case seems a straightforward example of the external criteria Zachar points at, since an external influence (the first interview) plays a role in how the subject decides to approach the concept of anhedonia. As such, in the practical kinds model reactivity may again be legitimate.

Nevertheless, as was the case in the harmful dysfunction model, there are some constraints on which classifications of mental disorders are appropriate in this model. Zachar argues that some practical kinds will be more 'reliable' (i.e. the kind which more often produces similar results under similar conditions), but in his practical model he does not explicitly state what might 'ground' such a more reliable kind. For that reason, I will devote the remainder of this section to the more modern descendant of Zachar's practical kinds model, the mechanistic property cluster (MPC) kinds model by Kenneth Kendler, Peter Zachar, and Carl Craver (Kendler et al., 2011). Analysing reactivity will turn out to be less straightforward for this model, as I will detail below.

Kendler et al.'s MPC kinds model is clear about what might ground a 'reliable' kind. Inspired by mechanistic theories in biology, this model attempts to fit disorders within networks of causal mechanisms called mechanistic property clusters or MPCs. Which concepts are acceptable depends on the causal network the disorder is a part of, i.e. on the causal mechanisms which affect the disorder. If researchers wish to use their concepts for prediction and intervention, their measurement must be closely tied to network analysis of the MPC. A reactive change to a characterization or representation that is inconsistent with these networks of causal mechanisms will therefore be problematic. There are some clear similarities here with the harmful dysfunction view, because whether or not one theorizes about the constraints on appropriate measurement of disorders in terms of mechanistic property clusters, part of the constraints will be biological (e.g. genetic or physiological) mechanisms. After all, these biological mechanisms are causally linked to the disorder and will as such be part of the complex causal network.

However, as was the case in the harmful dysfunction view, a straightforward 'essence' or set of boundary conditions for mental disorders is out of the question. Kendler et al. argue that the causality of mental disorders is so complex that MPCs do not refer to "simple, deterministic essences" (Kendler et al., 2011, 1146). They see MPC kinds, and thus mental disorders like depression, as akin to fuzzy sets, i.e. family resemblance sets with fuzzy boundaries. In Sect. 2, we have encountered this type of fuzzy phenomena under the name 'Ballung phenomena' (cf. Cartwright & Runhardt, 2014; Cartwright et al., 2017). Since depression is a Ballung phenomenon for Kendler et al., it is likely that more than one characterization of depression will be possible if we accept the MPC view of mental disorders, despite the close link to biological mechanisms.

Kendler et al.'s MPC analysis has been applied to the study of Major Depressive Disorder (MDD) in attempts to map the complex network of causal mechanisms which affect the development of MDD. A structured literature review by Andrea Wittenborn and colleagues (Wittenborn et al., 2016), in part inspired by the MPC model, has found the mechanisms affecting MDD include not only biological mechanisms, but also cognitive, social, and environmental drivers. To name just one example, dysfunctional behaviours related to MDD by an individual may lead to a lower quality of interpersonal relations (e.g. estrangement of the individual's significant other), leading to additional stress and negative affect, which again causes changes in the individual's behaviour. The complexity of the causal network, Wittenborn et al. conclude, makes the patient trajectories of individuals diagnosed with MDD highly idiosyncratic.

Does this mean that some beta and gamma reactivity is legitimate (does not undermine the accuracy of the measure) in the MPC model? Kendler et al. make no explicit mention of 'evaluative' aspects of characterizing mental disorders, which makes this less clear-cut than was the case in the harmful dysfunction view. Nevertheless, their MPC framework seems compatible with the existence of several different equally valid measures. Rather than being critical of social and personal influences on measurement, the MPC framework is intended as a strong recommendation to take causal mechanisms seriously when considering mental disorders. Although Kendel, Zachar, and Craver do not discuss the possibility of reactive changes in characterization and representation in the way I have described above, I believe their view is compatible with the idea that reactive beta and gamma change may be legitimate as long as the causal structure remains intact.

Whether beta and gamma change are legitimate will depend, however, on where on the scale of complexity a given mental disorder is situated. Kendler et al. describe this scale as running between simple disorders with a clear 'essence' and boundaries (akin to what the disease model postulates) on the one hand, and completely fuzzy kinds on the other (defined only by pragmatic or socially constructed criteria). If a given disorder turns out to sit on the 'constructed or practical kinds' end of the scale, beta and gamma change will be legitimate. If, alternatively, the mental disorder sits on the 'essentialist' end of the scale, such change will not be legitimate. This, respectively, mirrors the discussions on the human dysfunction (Sect. 4.2.3) and practical kinds view (above) on the one hand, and the discussion of the disease model in Sect. 4.2.1 on the other hand. We can only determine where on the scale depression sits through further careful causal mechanistic analysis, which is beyond the scope of this paper. However, studies like Wittenborn et al. (2016) clearly imply that depression may be too complex to be on the 'simple' end of this scale, which strengthens this article's general intuition that beta and gamma reactive effects will turn out to be legitimate for depressive disorders.

5 Conclusion and considerations for further research

Now that we have seen four different approaches to the relation between the phenomenon depression and the concept as described in the PHQ-9, let me sum up. In this paper, I have argued that we may think of reactivity as alpha, beta, or gamma

change, following Golembiewski et al.'s (1976) trichotomy; of these, beta and gamma change are more extreme. I have linked these types of change to, respectively, the representation and characterization stage of measurement in Cartwright and Runhardt's (2014) framework. I then showed that both beta and gamma change occur in measuring depression. I argued that under the most recent approaches to psychiatric classification, such changes may be legitimate, as long as they respect the biological or (more broadly) causal constraints these approaches point at. Legitimacy, here, meant specifically that such reactivity should not be corrected for by researchers, since it does not undermine the accuracy of the measure. Correcting for legitimate reactivity would ignore the autonomy of the research subject in controlling their characterization and representation of depressive phenomena.[23]

One possible reaction to the legitimacy of reactivity may be to see reactivity as somehow a tool for *improving* the validity of measurement. A practitioner may feel that, given that the subject is an authority on their affective states, reactivity will allow these symptoms and states to come to light better than before. This response, I would argue, is too quick. A baseline measure of a person's depressive symptoms will, in the best case scenario, be valid for the person (i.e. measure real mental states or physical properties) at that point in time. If, subsequently, a reactive change occurs, the next measurement may give a different result which, despite measuring with a different underlying representation or characterization, is as valid. So, we should not see reactive change as leading to an improvement in validity. Nor is it warranted, based on the arguments above alone, to infer that continued measurement should lead to more stable results. Whether reactive effects eventually 'converge' on some stable state or property can only be discovered through continuous investigation of reactive effects, but seems unlikely given the above-outlined complexity of the causal networks that disorders are a part of.

To finish this paper, let me now link my conclusion to a recent discussion in the literature, and use this link as a springboard for a question for further research. My conclusion is related to Jonathan Tsou's recent analysis of looping effects in psychiatry, due originally to Ian Hacking (Hacking, 1995, 1999, 2007). Tsou (2007, 2016, 2019), speaking specifically about whether depression is such a 'looping kind', argues that while *individuals* may experience reactive effects (changing in some way by being measured), this does not mean that meso-level social concepts like the concepts we use to classify depression are also changeable. He argues such concepts may well be stable because of the stable biological mechanisms that underly depressive symptoms (cf. Tsou, 2019, 190).

The latter two models I have focused on in this paper, the harmful dysfunction model and the practical kinds model, both imply that a stable set of biological mechanisms constrains measurement. Both models also leave room, however, for a change in characterization and representation, either because of a change in social values and meaning, or because of a change in other external, pragmatic factors.

My argument, that beta and gamma reactive change can be legitimate under both the harmful dysfunction model and practical kinds model, is closely related to an

[23] This should be distinguished from the traditional concept of *medical autonomy*, i.e. the right of subjects and patients to make informed decisions about their care (cf. The British Medical Association (BMA) 2020).

argument Tsou makes at the end of his 2007 paper. Here, he distinguishes two kinds of implications of looping:

(1) Weak implications of looping—Individuals' experiences and behaviours are altered in response to looping effects.
(2) Stronger implications of looping—Individuals' experiences and behaviours are altered in response to looping effects to the extent that the defining criteria for that classification change." (Tsou, 2007, 339–340)

This distinction is, I would argue, similar to a distinction between alpha change on the one hand and beta and gamma change on the other. After all, Tsou distinguishes two types of effects that reactivity can have; in the former situation, the classification does not change. He continues that in the second type of looping, reactive effects on a large number of individuals falling under the concept may suggest that the boundaries of the concept used to categorize them must be altered. He implies that such concepts would therefore become "unstable objects of knowledge" (Tsou, 2007, 341) and claims that Hacking has not shown that looping effects in actual psychological research and practice have such strong implications. Tsou believes that the biological process that underlies the characterization remains stable. In this paper, I have contributed to Hacking and Tsou's analysis of looping by further specifying what the stronger implications of looping may consist of.

The parallels between the Hacking and Tsou literature, and my argument in this paper, point at an important area for further research, viz. Tsou's concerns about the instability of knowledge. We must ask whether beta and gamma reactivity indeed lead to unstable knowledge, or whether the idea of legitimate reactivity that I argued for in this paper is compatible with a progressive psychometrics. In particular, we may ask to what extent an emphasis on legitimacy (as well as research subjects' authority in characterizing and representing phenomena) threatens researchers' inferences made on the basis of reactive measurement. I will finish by taking some first steps in answering this question.

As I have argued in the above, beta and gamma reactive change may be legitimate; therefore, trying to fix or avoid this change (as some methodologists have urged us to) would ignore the research subject's autonomy in characterizing and representing the phenomenon. This does not mean, however, that researchers should ignore such reactive change, never asking whether reactivity has occurred. I have argued briefly that finding out more about the changing characterization and representation of depression by individuals is a productive area of research. The first reason for this is intrinsic: there is value in learning more about the (often evaluative) choices individuals make in response to being measured. The second reason we should still investigate reactivity is instrumental: by asking whether reactive change has occurred, researchers can make stronger inferences. To illustrate, consider a PHQ measurement that has changed due to beta change, thereby creating an outlier in an otherwise stable timeseries of PHQ-scores for the patient.[24] Interpreting this

[24] Thanks to an anonymous reviewer for pointing out this illustrative example.

timeseries correctly is impossible without an investigation into the reactivity of the subject. In particular, we must avoid the hasty conclusion that the outlier is due to some other external influence it happens to be correlated with (e.g., a global health crisis just prior to the measurement of this outlier), rather than due to some legitimate reactive change.

While a more thorough investigation of how we can strengthen our inferences even in cases of legitimate reactivity must take place in further research, we can sum the above up as concrete advice for working clinical psychologists doing research on mental disorders like depression. Firstly, this paper has shown that under the harmful dysfunction and practical kinds models of mental disorders, we ought not to fix or avoid legitimate reactivity, since doing so would harm the autonomy of the research subject. Reactivity is not legitimate, however, when one can show that the subject's new concept or representation goes against the biological or causal constraints on the disorder. Second and finally, researchers ought to (continue to) investigate reactivity for both intrinsic and instrumental reasons outlined above.

Acknowledgements The author would like to thank Catherine Greene, the participants of the Reactivity and the Research Process workshop at the University of Bergen, March 5-6, 2020, as well as two anonymous reviewers for valuable critiques on earlier drafts of this paper.

Data availability Not applicable

Code availability Not applicable

Declarations

Conflicts of interest The author has no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ahmed, S., & Ring, L. (2008). Influence of response shift on evaluations of change in patient-reported outcomes. *Expert Review of Pharmacoeconomics & Outcomes Research, 8*, 479–489.

Alexandrova, A. (2018). Can the science of well-being be objective? *The British Journal for the Philosophy of Science, 69*(2), 421–445.

APA. (1980). *Diagnostic and statistical manual of mental disorders* (Third Edition: DSMIII). American Psychiatric Association.

APA. (2000). *Diagnostic and statistical manual of mental disorders* (Fourth Edition Text Revision: DSM-IV-TR). American Psychiatric Association.

APA. (2011). Patient Health Questionnaire (PHQ-9 and PHQ-2). *American Psychiatric Association.* http://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/patient-health. Accessed 24 June 2020.

APA. (2013). *Diagnostic and statistical manual of mental disorders* (Fifth Edition: DSM-5). American Psychiatric Association.

BMA. (2020). "Autonomy or self-determination as a medical student." https://www.bma.org.uk/advice-and-support/ethics/medical-students/ethics-toolkit-for-medical-students/autonomy-or-self-determination. Accessed 16 Feb 2021.

Cartwright, N., & Runhardt, R. W. (2014). Measurement. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (pp. 265–287). Oxford University Press.

Cartwright, N., Bradburn, N., & Fuller, J. (2017). "A theory of measurement." In L. McClimans (Ed.), *Measurement in medicine: philosophical essays on assessment and evaluation.* Rowman & Littlefield.

Choquette, K. A., & Hesselbrock, M. N. (1987). Effects of retesting with the Beck and Zung depression scales in alcoholics. *Alcohol and Alcoholism, 22*, 277–283.

Cooper, R. (2004). *Classifying madness: A philosophical examination of the diagnostic and statistical manual of mental disorders*. Springer.

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology, 15*, 453–468.

French, D. P., & Sutton, S. (2011). Does measuring people change them? *The Psychologist, 24*, 272–274.

Gibbons, F. X. (1999). Social comparison as a mediator of response shift. *Social Science & Medicine, 48*, 1517–1530.

Glynn, L. G., Hayes, P. S., Casey, M., Glynn, F., Alvarez-Iglesias, A., Newell, J., ÓLaighin, G., Heaney, D., O'Donnell, M., & Murphy, A. W. (2014). Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. *British Journal of General Practice, 64*, e384.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science, 12*, 133–157.

Greene, C. (2019). Nomadic concepts, variable choice, and the social sciences. *Philosophy of the Social Sciences, 50*, 3–22.

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press.

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Hacking, I. (2007). Kinds of people: Moving targets. In *Proceedings of the British Academy, Volume 151, 2006 Lectures*. Vol. 151.

Haybron, D. M. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Noûs, 41*, 394–428.

Horwitz, A. V. (2002). *Creating mental illness*. University of Chicago Press.

Horwitz, A. V. (2014). The social functions of natural kinds: The case of major depression. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying psychopathology: Mental kinds and natural kinds* (pp. 209–229). MIT Press.

Horwitz, A. V., & Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford University Press.

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the patient health questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine, 21*, 547–552.

Jiménez-Buedo, M. (2021). Reactivity in social scientific experiments: What is it and how is it different (and worse) than a placebo effect? *European Journal for Philosophy of Science, 11*(42), 1–22.

Johar, O., & Sackett, A. M. (2018). The self-contaminating nature of repeated reports of negative emotions. *Basic and Applied Social Psychology, 40*, 293–307.

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine, 41*, 1143–1150.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*, 509–515.

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., & Rathouz, P. J. (2008). Self-reported and measured sleep duration: How similar are they? *Epidemiology, 19*, 838–845.

Little, D. (1993). On the scope and limits of generalizations in the social sciences. *Synthese, 97*, 183–207.

Marsay, C., Manderson, L., & Subramaney, U. (2018). Changes in mood after screening for antenatal anxiety and depression. *Journal of Reproductive and Infant Psychology, 36*, 347–362. Routledge.

Norman, P., & Parker, S. (1996). The Interpretation of change in verbal reports: Implications for health psychology. *Psychology and Health, 11*, 301–314.

Sajobi, T. T., Brahmbatt, R., Lix, L. M., Zumbo, B. D., & Sawatzky, R. (2018). Scoping review of response shift methods: Current reporting practices and recommendations. *Quality of Life Research, 27*, 1133–1146.

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine, 48*, 1531–1548.

Sharpe, P. J., & Gilbert, D. G. (1998). Effects of repeated administration of the beck depression inventory and other measures of negative mood states. *Personality and Individual Differences, 24*, 457–463.

Smarr, K. L. (2003). Measures of depression and depressive symptoms: The Beck Depression Inventory (BDI), Center for Epidemiological Studies-Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Primary Care Evaluation of Mental Disorders-Mood Module (PRIME-MD). *Arthritis Care & Research, 49*, S134–S146.

Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine, 48*, 1507–1515.

Tsou, J. Y. (2007). Hacking on the looping effects of psychiatric classifications: What is an interactive and indifferent kind? *International Studies in the Philosophy of Science, 21*, 329–344.

Tsou, J. Y. (2016). Natural kinds, psychiatric classification, and the history of the DSM. *History of Psychiatry, 27*, 406–424.

Tsou, J. Y. (2019). Philosophy of science, psychiatric classification, and the DSM. In S. Tekin & R. Bluhm (Eds.), *The bloomsbury companion to philosophy of psychiatry* (pp. 177–196). Bloomsbury.

VanderZee, K. I., Buunk, B. P., & Sanderman, R. (1995). Social comparison as a mediator between health problems and subjective health evaluations. *British Journal of Social Psychology, 34*, 53–65.

Wakefield, J. C. (1992). Disorder as harmful dysfunction: A conceptual critique of DSM-III-R's definition of mental disorder. *Psychological Review, 99*, 232–247.

Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.

Wittenborn, A., Rahmandad, H., Rick, J., & Hosseinichimeh, N. (2016). Depression as a systemic syndrome: Mapping the feedback loops of major depressive disorder. *Psychological Medicine, 46*(3), 551–562.

Wood, J. V., Taylor, S. E., & Lichtman, R. R. (1985). Social Comparison in Adjustment to Breast Cancer. *Journal of Personality and Social Psychology, 49*, 1169–1183.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Zachar, P. (2000). Psychiatric disorders are not natural kinds. *Philosophy, Psychiatry, and Psychology, 7*, 167–182.

Zachar, P. (2002). The practical kinds model as a pragmatist theory of classification. *Philosophy, Psychiatry, and Psychology, 9*, 219–227.

Zachar, P. (2014). Beyond natural kinds: Toward a "Relevant" "Scientific" taxonomy in psychiatry. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying psychopathology: Mental kinds and natural kinds* (pp. 75–104). MIT Press.