European Journal for Philosophy of Science (2021) 11:104
https://doi.org/10.1007/s13194-021-00424-y

PAPER IN PHILOSOPHY OF SCIENCE IN PRACTICE

From measurement to classificatory practice: improving psychiatric classification independently of the opposition between symptom-based and causal approaches

Alessandra Basso[1]

Received: 12 January 2021 / Accepted: 13 October 2021

Abstract
The article advances a new way of thinking about classifications in general and the classification of mental disorders in particular. By applying insights from measurement practice to the context of classification, I defend a notion of epistemic accuracy that allows one to evaluate and improve classifications by comparing different classifying methods to each other. Progress in classification arises from the mutual development of classification systems and classifying methods. Based on this notion of accuracy, the article illustrates with an example how psychiatric classifications can be improved via circumscribed comparisons of different perspectives on mental disorders, without relying on complete models of their complex aetiology. When applying this strategy, the traditional opposition between symptom-based and causal approaches is of little consequence for making progress in the epistemic accuracy of psychiatric classification.

Keywords Philosophy of measurement · Epistemology of classification · Psychiatric representation · Measurement in science · Psychiatric classification · Accuracy of classification · Classification · Classification of mental disorders

1 Introduction

Established measurements are considered reliable sources of knowledge throughout the sciences as well as in lay contexts. In medicine, measurement has recently gained even sharper relevance due to the rise of the evidence-based movement, which encourages the search for accurate ways to diagnose and classify diseases for clinical, statistical and research purposes (McClimans, 2017; Sackett, 1997, 2000; Guyatt, 1992). However, today as in the past, cutting-edge measurement practice is

✉ Alessandra Basso
alessandra.basso@helsinki.fi

[1] Practical philosophy, University of Helsinki, Helsinki, Finland

Published online: 02 November 2021

not a straightforward matter: before a measurement can be taken as granted as a knowledge-producing activity, scientists face the challenge of developing new measurement procedures and improving their accuracy without independent epistemic access to what is being measured. In this process, identifying what is being measured and discovering how to measure it are not independent tasks, but are instead thought to coevolve.

This paper provides a measurement perspective on the classification of mental disorders based on the idea that classifications can be considered cases of measurement in which the outcomes are nominal rather than quantitative (Stevens, 1946; Tal, 2020). Psychiatric classifications are discussed as forms of measurement within the relevant disciplines, such as in psychometric theories about the relation between symptoms and constructs (Borsboom, 2008; Molenaar, 2004; Hand, 2004; McClimans, 2017). Here, however, I rely on a recent philosophical literature on measurement that emphasizes the mutual development of theoretical and procedural aspects of measurement, and in particular on the so called model-based accounts that address measurement assessment practice and the notion of accuracy that emerges from it (Chang, 2004; van Fraassen, 2008; Tal, 2016, 2017, 2019; Frigerio et al., 2010; Mari, 2005).

Although this literature refers mainly to the measurement of physical quantities, it has broader applicability. Insights from these philosophical works have been applied successfully to the social sciences, as well as in psychology and psychiatry (Alexandrova, 2016; Boumans, 2015; Basso, 2017; Bringmann & Eronen, 2016; Kendler, 2012a; McClimans, 2017; Wilson, 2013; Mari & Wilson, 2014; McClimans et al., 2017). I suggest extending its application to classifications in general and psychiatric classifications in particular. This framework proves useful because it allows us to see that the debate about psychiatric taxonomy faces problems that are common to all new measurements procedures, and it provides a novel way to address these problems.

The widespread dissatisfaction with the current official classification of mental disorders led to heated controversies about the future of psychiatric taxonomy (Demazeux & Singy, 2015; Kendler & Parnas, 2017; Kendler, 2012b; Kendler & First, 2010; Zachar et al., 2014; Cooper, 2018; Tabb, 2017). The debate revolves around the opposition between classifications based on the patients' symptoms and classifications based on the underlying causes of mental disorders. The modern, symptom-based classifications, which are purportedly atheoretical, are criticized for being inadequate to clinical, research and public health purposes: they are held responsible for the limited efficacy in the treatment and prevention of psychiatric disorders; they are thought to limit research progress into the causes of mental disorders; and they are taken to be the reason for the lack of improvement in the mortality rate associated with these diseases (Cuthbert & Insel, 2013; Cuthbert, 2015). As a result, many in the field recommend getting theoretical commitments back into the picture (Kendler, 2005, 2012a; Murphy, 2006; Hyman, 2007; Kupfer & Regier, 2011; Follette & Houts, 1996; Tsou, 2015).

What is at stake is the inclusion in the classification of the emerging knowledge about the causal processes underlying mental disorders. Mental disorders arise from multiple biological, psychological and environmental factors that are

best investigated from a variety of disciplinary perspectives, and interact with one another in complex ways. Symptom-based classifications appear too simplistic when confronted with this complexity, and are deemed inadequate to capture what is important concerning treatment and prevention. Based on the idea that knowing the causes of mental disorders would allow for better treatment and prevention, some psychiatrists recommend moving to a classification able to reflect the complex aetiology of mental disorders (Murphy, 2006; Kendler, 2012b; Cuthbert & Insel, 2013).

My motivations for adopting a measurement perspective in addressing this debate are the following: 1) it clarifies that the problems of psychiatric classification are common to all new measurements; 2) it allows to interpret the psychiatric debate as revolving around two accounts of measurement, conventionalism and realism, which are both problematic and do not offer strategies for evaluating and improving classificatory practice; and 3) it suggests an alternative approach, based on a metaphysically neutral notion of epistemic accuracy that can be used to evaluate and improve the classifications.

First, one of the lessons learned from recent historical and philosophical works on measurement is that the theory of the phenomenon and its measurement coevolve (Chang, 2004; van Fraassen, 2008). To quote Bas van Fraassen: "The questions *What counts as a measurement of (physical quantity) X?* and *What is (that physical quantity) X?* cannot be answered independently of each other" (van Fraassen, 2008, p. 116, original emphasis). It is not surprizing, therefore, that the taxonomic debates revolve around two main questions: how to conceptualize mental disorders and what is the most appropriate epistemological basis for classifying them. Redefinitions of the phenomena and of the way to measure or classify them are common in the development of scientific practice, and for good reasons: it is their mutual refinement that allows for progress.

The second advantage of adopting a measurement perspective is that it allows to interpret the debate as revolving around two accounts of measurement: conventionalism and realism, which reflect different views on what is the nature of measurement and what conditions make it reliable. Symptom-based classifications can be seen as relying on a conventionalist view of measurement, which focuses on providing an agreed-upon set of diagnostic rules, so as to improve the repeatability of diagnoses across different clinicians. Causal classifications, instead, reflect a realist view, which prioritizes the aim of capturing the real phenomenon of interest, so that the results convey information about what is being measured. When it comes to evaluate and improve the accuracy of classificatory practice, however, both accounts are problematic. Symptom-based classifications rely on an operational notion of accuracy, which depends on the compliance to agreed-upon definitions. This notion of accuracy is problematic because it is exposed to the objection of being partially arbitrary and lacks a clear justification of why these definitions should be standard. Causal classifications, on the other hand, appear to be based on a metaphysical notion of accuracy, which depends on having access to how mental phenomena are split and lumped independently of our classifications. This notion of accuracy is problematic because true classifications are unknowable, and hence the accuracy of classification remains undetermined. At this point, one might question whether it is possible at all to improve psychiatric classification.

A third advantage of adopting a measurement perspective is that it suggests that improving the accuracy of classifications is a feasible goal if one adopts the epistemic notion of accuracy derived from measurement assessment practice. This notion of accuracy depends on the coherence between the abstract classification terms and the classificatory procedure, and it is formulated by introducing a distinction between *classification system* and *classifying method*: the former refers to how phenomena are ideally split and lumped according to an underlying organizing principle; the latter indicates the concrete procedure for assigning single cases to classes. As theory and measurement coevolve, classification systems and classifying methods can also be seen as being part of a process in which the two mutually influence each other. Recent research efforts into the aetiology of psychiatric disorders can be interpreted as providing opportunities to improve epistemic accuracy by comparing different classificatory methods to each other. Scientists compare disciplinary perspectives on specific mental disorders, based on 'local' representations of the interaction between different causal factors. In this process, classification systems and classifying methods coevolve by mutually refining each other. The success of this strategy does not depend entirely on whether mental disorders are initially classified in terms of symptoms or causes. This suggests that perhaps the debates over psychiatric taxonomy have been overly concerned with the opposition between these alternative approaches, because this opposition is of little consequence for making progress in the epistemic accuracy of psychiatric classification.

The paper proceeds as follows: Section 2 invites a reinterpretation of the debate about psychiatric classification and argues that it can be seen as revolving around two underling views of measurement: conventionalism and realism. Section 3 draws on insights from the recent philosophical literature on measurement for providing an epistemic account of measurement accuracy to be applied in this context. Section 4 argues that symptom-based and causal approaches are based on problematic notions of accuracy, and it suggests moving to an epistemic notion instead. In Section 5, I illustrate with an example how it is possible to improve the classification of mental disorders by comparing different disciplinary perspectives on specific mental phenomena. Section 6 concludes by emphasizing that successful improvements of psychiatric classifications are relatively independent of the opposition between symptom-based and causal approaches.

2 Rethinking the debate between symptom-based and causal approaches

The debate about psychiatric classification revolves around the opposition between modern classifications based on symptoms and the proposals to classify mental disorders according to their causes. Symptom-based classifications are also said to be *descriptive* in contrast to classifications driven by *theoretical* commitments about the causes of mental disorders. Symptom-based classifications are said to be descriptive and atheoretical because they appeal to the empirical observation of correlations among symptoms, and avoid being driven by theoretical commitments about the possible causes of the disorders. This characterization, however, does not fully

capture the distinction between the two approaches, and the historical motivations for the emergence of symptom-based classifications.

By looking at the history of psychiatric taxonomy, indeed, it emerges that the move from 'theoretical' to 'descriptive' classifications was not meant to dispense with theoretical commitments altogether (which is not possible anyway), but was rather motivated by the aim of standardizing diagnostic practice so as to facilitate comparisons and scientific communication, and improve inter-rater reliability. Symptom-based classifications originated in the late 1950s from a convergence of factors, both internal and external to psychiatry (Wilson, 1993). The early, so called 'theoretical' classifications of mental disorders were not well received by the World Health Organization (WHO) member states, in contrast to the other chapters of the *International Classification of Diseases* (ICD-6 and ICD-7), which were readily adopted. So in 1959, psychiatrist Erwin Stengel was commissioned by WHO to look into ways of improving the classification of mental disorders. His analysis, and the debates that rose within WHO and the American Psychiatric Association (APA), brought to light not only a multiplicity of diagnostic terms employed around the world, but also a thorough disagreement among psychiatrists about their theoretical perspectives (Blashfield et al., 2014).

Although Stengel's (1959) final report to WHO did not contain a clear recommendation to move towards a symptom-based classification, the suggestion emerged that definitions based on symptoms, devoid of theoretical commitments and ambiguous terms, would be best in order to facilitate wide adoption by member states, thereby improving the comparability of findings and their use for public health purposes.[1] This is the idea that shaped the revision of the classifications of mental disorders. The ICD-8 (WHO, 1967, 1974) and the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III, APA, 1980) became the first predominantly symptom-based classifications of mental disorders, which are said to have marked a paradigm shift towards a descriptive classification (Fulford & Sartorius, 2009; Blashfield et al., 2014; Clark et al., 2017; Compton & Guze, 1995; Wilson, 1993).[2]

As stated in the introduction, DSM-III is explicitly devoid of theoretical commitments regarding the origins of psychiatric disorders in order to facilitate adoption of the manual by psychiatrists of different theoretical orientations:

> The approach taken in DSM-III is atheoretical with regard to etiology or pathophysiological process except for those disorders for which this is well established and therefore included in the definition of the disorder. Undoubtedly, with time, some of the disorders of unknown etiology will be

[1] See Fulford and Sartorius (2009) for a historical reconstruction.

[2] Unlike their predecessors, symptom-based classifications were widely adopted by nearly all WHO member states and were well received by clinicians and institutions across mental health disciplines. This approach became increasingly dominant and was received with excitement both in mental health disciplines (where it gave rise to an expansion of empirical research aimed at measuring the prevalence and course of psychiatric diseases) and by the wider public. Subsequent revisions of the manuals maintained the same conceptual approach.

> found to have specified biological etiologies, others to have specific psychological causes and still others to result mainly from a particular interplay of psychological, social and biological factors.
> The major justification for the generally atheoretical approach taken in DSM-III with regard to etiology is that the inclusion of etiological variables would be an obstacle to use of the manual by clinicians of varying theoretical orientations, since it would not be possible to present all reasonable etiological theories for each disorder. (APA, 1980, p. 7)

The classification is neutral with respect to different theoretical perspectives about the aetiology of mental disorders and minimizes the use of terms with disputable meanings or those susceptible to divergent interpretations (Lalumera, 2016). Ideally, this allows for objective diagnoses, objective in the sense of being independent of the clinician's theoretical inclination, and thereby improves inter-clinician reliability (that is, the amount of variance among psychiatrists' diagnostic judgments). In substance, the shift to symptom-based classification was motivated by the aim of standardizing the language and the diagnostic practice among practitioners, so as to foster homogeneous diagnoses, facilitate scientific communication and improve the accumulation of knowledge. Multiple theoretical approaches coexist in psychiatry, but diagnoses should not be influenced by theoretical disagreement. The causal approaches to psychiatric classification, however, do not mean to change this: they, too, aim for objective diagnoses in this sense. Indeed, the attempt to establish biological markers for identifying mental disorders can be seen as being motivated by the aim of reducing what is sometimes perceived as an over-reliance on clinical judgment.

The opposition between causal and symptom-based approaches therefore, is not simply about appealing to theory versus relying on empirical observation. Indeed, causal classifications, too, are based on empirical observation, such as brain scan images, electrocardiograms, and genome-wide association studies. More importantly, moreover, while symptom-based classifications avoid theoretical commitment specifically about the aetiology of mental disorders, it would seem that any classification, no matter how empirically based, necessarily presumes some background theory. After all, empirical evidence alone does not determine how to split and lump clusters of symptoms. The problem is that, in the case of symptom-based classifications, it is not clear what theoretical principles drive the choice of diagnostic criteria. Historically, symptom-based classifications prioritised the agreement on a standard set of diagnostic rules to bypass the thorough theoretical disagreement about the causes of mental disorders. In this sense, the atheoretical stance of symptom-based classifications was a means to promote their acceptance, but it also raised questions about the justifications of the diagnostic criteria (Follette & Houts, 1996; Wilson, 1993). The diagnostic manuals avoid engaging with theoretical issues (about aetiology or something else) for guiding the choice of diagnostic rules. As a consequence, in the absence of explicit discussion about the underlying theoretical and empirical principles, the choice of diagnostic criteria appears to be based mostly on conventions (Wilson, 1993).

This leads to interpret the debate about psychiatric classification as reflecting two underlying accounts of measurement, conventionalism and realism. Symptom-based classifications can be interpreted as being based on a broadly conceived conventionalist view, which emphasizes that there is a conventional element to the use of classification terms like depression or schizophrenia. On this view, the classification of individuals (what give empirical content to class-terms?) is performed according to an agreed-upon set of rules, rather than by empirically and theoretically justified criteria. Symptom-based classifications standardize the definitions of mental disorders by prescribing a standard way in which these concepts are to be applied to particular individuals. What makes it conventionalist is that these standard rules appear to involve nontrivial choices made by humans, which do not need to satisfy theoretical constraints (Ellis, 1966).[3]

The problem with this approach is that the standard diagnostic rules appear to lack theoretical justification, and are therefore exposed to the objection of being arbitrary. Diagnoses based on standard rules are repeatable because they can be performed in the same way by different clinicians. But the choice of standards is partially based on conventions, and hence can raise questions about why these particular rules should be standard.[4] Moreover, if the choice of standards is fully conventional, it is not clear whether and how it can be improved: what criterion could be used to identify more accurate rules? (Tal, 2020).

Causal classifications, in contrast, are based on a realist approach, according to which class-terms refer to how phenomena are split and lumped independently of the conventions and the theoretical orientation of the person who classify them. In this view, the classification of individuals is an approximation of true taxonomies, which provides the best explanation for the usefulness and the repeatability of diagnostic practice (Byerly & Lazara, 1973; Swoyer, 1987). For instance, different diagnoses give similar results because they are sensitive to the same facts. Realists explain the accuracy of classificatory practice, and its improvement, in terms of the distance from real taxonomies. The problem with this view is that true taxonomies are unknowable (even admitting there are such things), and hence the accuracy of diagnostic practice remains undetermined.

Realism and conventionalism are in principle compatible, because it is conceivable that *if* the choices of standards are correct, the class-terms succeed in referring

[3] Symptom-based classifications identify mental disorders by means of their diagnostic criteria, but the definitions are not strictly operational in the sense proposed by Bridgman (1927), because they are not based on a tautological identification between categories and diagnostic criteria. Instead, the classifications admit different procedures for monitoring symptoms (a variety of questionnaires and interviews), and do not exclude other methods if they are available.

[4] A different question is whether, despite the avoidance of explicit engagement with theoretical issues, the symptom-based approach could be justified on the basis of a given theory. One reviewer suggested that Classical Test Theory could be used to defend symptom-based classifications from the accusation of lacking theoretical justification. This is an interesting idea that calls for further examination (cf. McClimans et al., 2017). Such defence, however, might still be exposed to the objection that clusters of symptoms are only superficial correlations, which does not allow to identify classes that are relevantly homogeneous, for instance by having similar phenotypic and genotypic characteristics.

to real taxonomies of phenomena (Tal, 2020).[5] Indeed, from a realist, error-based perspective, these approaches to psychopathology can be seen as prioritizing complementary aspects of measurement reliability. Symptom-based approaches put the accent on standardizing classificatory practice in order to improve *reliability*. Causal approaches, instead, focus on the content of class-terms and their ability to capture the phenomenon of interest, which can be referred to as *validity*.

As noted above, however, both approaches are problematic, especially if they are seen are being in opposition to each other. The recent philosophical literature on measurement provides an alternative view, which combines issues about the practicalities of classification and about their theoretical content. On this view, to gain empirical significance, the abstract classification terms have to be linked, or 'coordinated' with procedures that enable to assign particulars to classes. The so-called 'problem of coordination' is that, prior to the development of an accepted classificatory method, there is no evidence to confirm the rule for assigning particulars to the classes of interest. Insights from the recent literature on measurement offer a perspective in which the coherence between class-terms and classifying procedure is an alternative solution to this problem, which does not appeal to an independent truth, and neither aims only at the consistency of results over repetition.

3 Insights from the epistemology of measurement

Recent literature in the philosophy of measurement, most notably including the works of Hasok Chang and Bas van Fraassen, looked with new emphasis at the problem coordination between theoretical parameters and their empirical content, and put forward a novel and largely coherentist solution to this problem, based on the historical development of measurement and its epistemic justification (Chang, 2004; van Fraassen, 2008). In contrast to earlier foundational approaches, which aimed at finding an empirical and theory-independent relation between measurement outcomes and targeted parameters, these authors argue that the epistemic value of measurement can be understood only by looking at the mutual development of theory and measurement procedure.[6]

In his investigations on the history of thermometry, Chang emphasizes that any attempt to confirm the results of thermometers on the basis of evidence alone leads to circularity, because, prior to the construction of an accepted measurement procedure, there is no evidence to confirm the law according to which inferences are made from observed changes in volume to changes in temperature. As a consequence, issues such as whether the changes in volume of a certain substance are quantitatively proportional to changes in temperature remain underdetermined by evidence (Chang, 2004, Ch. 2). Instead of trying to avoid the circularity, Chang provides an interpretation to show that it is not vicious. According to Chang, the historical

[5] Stronger versions of operationalism and conventionalism, according to which there is no fact of the matter about what is the empirical content of class-terms, instead, are incompatible with realism.

[6] See Tal (2020) for a review of foundational approaches and a comparison with recent works.

progress of thermometry involved back-and-forth relations between empirical interventions and theory developments. On the one hand, the construction and testing of thermometers required underlying theoretical assumptions that could only be provisional (e.g., the linear expansion of thermometric substances) or more widely accepted thanks to confirmation coming from other fields (e.g., the law of thermal expansion of gases). On the other hand, the gathered empirical evidence could lead scientists to amend and refine theory and its concepts, e.g. to cast doubt on the assumption of linear expansion of thermometric substances. Each step improved on the previous conceptualization of the quantity and standardization of its measurement, and refined the coherence among them.

In the absence of independent epistemic access to the quantity of interest, measurement can be justified only in the light of the mutual development of theory and measurement procedure. Van Fraassen maintains that measurement outcomes convey information about the quantity under measurement only relative to a theoretical interpretation of the physical interaction between the instrument and the object, which grounds the assignment of values to the respective quantities. It is only from the perspective offered by the theoretical interpretation of the measurement at hand that we can see how measurement assigns a value for what is measured.

In a recent series of papers, Tal further develops this idea and puts forward a model-based account of measurement that addresses not only the conundrum of underdetermination discussed above, but also other issues that are central to contemporary metrology, the science of measurement, such as the evaluation of uncertainty and the correction of systematic errors (Tal, 2019, 2016, 2011). Tal distinguishes between instrument *indications*, such as the position of the fluid inside a thermometer or the location of the pointer on a balance, and measurement *outcomes*, which are knowledge claims about the quantity being measured, like 'the temperature of x is 23°C' and 'the weight of y is 1.2 kg' (see also Frigerio et al., 2010). According to model-based accounts, measurement involves making inferences from indications to outcomes, but these inferences cannot be made on the basis of indications alone. Why? On the one hand, it is because of the same problem noticed by Chang and van Fraassen: by looking at indications alone, one cannot find out what is the inferential rule for deriving claims about the quantity under measurement. On the other hand, Tal notices that instrument indications are subjected to the idiosyncrasies of the concrete measurement process, such as the interferences coming from the environment, the operator and the particular features of the instrument. For example, in the measurement of temperature, indications are influenced by changes in the volume of the thermometer glass. Measurement outcomes, instead, are expected to be invariant to these interferences, so as to yield knowledge that is independent from interfering factors. By looking at indications alone, one cannot distinguish between variations that are due to the influence of interfering factors from variations due to changes in the quantity under measurement.

So, the question arises of how measurement can yield justified knowledge claims. Like Chang's and van Fraassen's, Tal's answer to this question also involves a reference to theory. Tal's account, however, is more specific on what is the theoretical background that is involved in the understanding and justification of measurement. In his view, measurement presupposes a hypothetical

representation of the measurement process, that is, a model that represents the quantity under measurement and the measurement instrument in the ideal situation in which interfering factors are absent or controlled for.

Both of the above problems find an answer by reference to this model. By making assumptions on how the measurement process would be in isolation from all interfering factors, this model:

a) provides an interpretation of the measurement process in terms that relate the instrument indications to the quantity of interest, thereby justifying how measurement can assign values for what is measured. In other words, this representation provides the inferential rule for deriving knowledge claims from instrument indications;
b) provides the theoretical basis to detect interfering factors and possibly control for their effects. According to model-based accounts, measurement errors are evaluated by the distance between a given outcome and the value that would be expected in the absence of interfering factors, and hence are revealed as the discrepancies between the actual outcomes and the model predictions (Tal, 2019, p. 871; Tal, 2017; Basso, 2017). On this view, therefore, the accuracy of measurement has to do with the fit between the ideal situation and the actual process, and depends on the process as well as on how it is represented: errors are defined as the distance between what is observed and what would be expected in the ideal situation described by the model of how the measurement works (Tal, 2017). This notion of accuracy can account for the evaluation of uncertainty and the correction of systematic errors in metrology (JCGM, 2012).

In the absence of independent epistemic access to the quantity under measurement, scientists have developed strategies to detect, distribute and correct errors by comparing outcomes to each other. The comparison of different measurements of the same quantity is central to measurement assessment practice, and it is common across the natural and the social sciences (Staley, 2020; Basso, 2017). The consistency of outcomes across different measurements of the same quantity is meant to ensure that the outcomes can be ascribed to what is measured, rather than to some artefact of the measuring instrument, the environment or the model. Tal (2019) emphasizes that this comparison is meaningful only if the measurements are modelled in terms of the same quantity of interest. Conditional on this judgment, agreement can be taken as a sign of accuracy, and discrepancies can be interpreted as pointing to undetected errors: by comparing measurement outcomes in the light of their respective models, scientists can detect errors due to different interfering factors. This involves both manipulating the model and intervening in the process: improving the accuracy of measurement might require altering the process, but might also be done without physical interventions, by adjusting or modifying the representation (Tal, 2019). Measurements are deemed accurate if their outcomes agree within their respective uncertainty intervals (Tal, 2011; Basso, 2017).

Continuing the example from the history of thermometry, scientists detected the interfering effect of the different glasses' rates of expansions by observing

the disagreement between thermometers. Once the outcomes are corrected so as to account for this interfering factor, the thermometers are made to agree without physical interventions (Chang, 2004). Similarly, a questionnaire and an interview for diagnosing migraine can be found to disagree because the former, but not the latter, tend to underestimate the duration of headache due to the subjects taking medication or sleeping through it (Basso, 2017).[7] Once the error is corrected for, the two diagnostic methods are made to agree with one another within their respective uncertainty levels. Since the correction of systematic errors may lead to an increase in uncertainty, accuracy is improved when the errors are corrected with a relatively small increase of uncertainty (Tal, 2019).

Discrepancies, however, do not need to be resolved in favour of error detection. A disagreement between measurement outcomes can be interpreted as pointing to an undetected error, or as revealing that the instruments measure different quantities. According to Tal, this is because, "under the model-based account, the distribution of systematic errors and the individuation of quantities are but two sides of the same epistemic coin. Which side of the coin the scientific community should fall on when resolving the next discrepancy depends on the particular history of its theoretical and methodological development. There is no *context-free* way to decide" (Tal, 2019, original emphasis).

4 The accuracy of classifications

The most common way to evaluate the accuracy of diagnostic methods is in terms of sensitivity and specificity, where sensitivity is defined as the probability of a positive diagnosis when the condition is present, and specificity is defined as the probability of a negative diagnosis when the condition is absent. In medicine, there are various ways to combine these probabilities to evaluate diagnostic practice, as well as alternative strategies to choose the diagnostic thresholds so as to minimize the costs of misclassification (Hand, 2004, p. 198-201). This method of evaluation applies to diagnostic methods that assign people to two classes: diseased and non-diseased.

The evaluation of psychiatric taxonomies, however, is more complex, not only because the classification has more than two classes, but also because the demarcation between classes is itself a matter of debate. As a consequence, the accuracy of a classification can have two different meanings, which raise different epistemic problems. First, the accuracy of a *classification system* refers to whether it reflects the desired underlying organizing principle. For instance, the organizing principle can be metaphysical – when it refers to an independent reality; or it can come from theory – when it refers to theoretical explanations (such as the aetiology of phenomena); or it can be based on a practical goal – such as maximizing treatment success.

[7] Construct validation also relies on the comparison of similar or related measurements to evaluate the validity of psychological constructs (Campbell & Fiske, 1959). This method is widely used in psychology and psychiatry to evaluate the validity of questionnaires, and it bears some similarity to the evaluation of accuracy in metrology (Wilson, 2013).

In evaluating this notion of accuracy, scientists face the problem of how one can tell that a classification matches the desired principle of organization.

Second, one can talk about the accuracy of the *classifying method*: even if the classification system is perfect (that is, it perfectly reflects the desired organizing principle, and assuming we have a way to know that it does), the classifying method could still fail, resulting in misclassifying particulars. In this case, the accuracy refers to the classification of the single cases: it evaluates whether the classifier assigns individuals to the right class (where 'right' means in accordance with the desired organizing principle). In evaluating this notion of accuracy, the question arises of how one can evaluate if the method classifies individuals correctly.

Ideally, a classification should be accurate according to both of these meanings. In psychiatry, however, both the classification system and the classifying method are at the centre of heated debates, so we face both epistemic problems at the same time and this generates deep epistemic uncertainty. On the one hand, there is debate about which should be the organizing principle for classifying psychiatric disorders. This depends not only on the debate between symptom-based and causal approaches, but also on the coexistence of various disciplinary perspectives on how to split and lump mental phenomena (Stinson, 2016; Sullivan, 2008; Marchionni manuscript). Different disciplines use independent sources of evidence and have different criteria for carving the causal field. As a consequence, field-specific representations of mental phenomena can be incompatible, or even conflicting, and still be regarded as good models both within and outside their disciplines. For instance, Stinson highlights that cognitive models are incompatible in several ways with current neuroscientific models, and yet they are not rejected or regarded as inaccurate. On the other hand, even if we agreed upon one classification system, we would need a gold standard for evaluating the accuracy of classifying methods. Without an external reference against which to assess sensitivity and specificity, we cannot estimate the error rates. In other areas of medicine, there are such external references: a new method for diagnosing breast cancer, for instance, could be evaluated against a biopsy. In psychiatry, there are no such external standards.

In addressing this situation, one of the advantages of taking a measurement perspective is that it allows us to realize that this is a general problem with new measurement procedures, which are not yet established in scientific practice. As theory and practice of measurement co-evolve in the development of measurement, taxonomic systems and classification of particulars can also be part of a process in which the two co-evolve and mutually influence each other. This framework suggests that it might be possible to make step-by-step progress by means of mutual influence between the two. A better classification system can refine the classification of particulars, which in turn can be tested more precisely, and thereby allow for further progress. Insights from measurement practice can help in finding a notion of accuracy that enables one to deal with the problems that arise when evaluating the accuracy of classifications, and can suggest ways to improve them.

The current debates in psychiatry are polarized by two opposed notions of accuracy that are both problematic and do not provide tools to address the evaluation and improvement of the classification. Symptom-based classifications appear to be based on an operational notion of accuracy. They standardize the definitions of class-terms

so that all practitioners can talk the same language. Section 2 clarified that these classification systems aim to maximize the repeatability of diagnoses, and do so by defining the disorders in terms of conventional indications of how to identify them. In this context, accuracy depends on the compliance to agreed-upon definitions, but this notion of accuracy is problematic because it involves an element of conventionality and lacks a clear justification of why these definitions should be standard. Moreover, symptom-based classifications standardize the diagnostic practice, but there is no error-correcting mechanism for improving the classifying method. Ad hoc adjustments are introduced if evidence highlights problems, but without identifying and correcting the source of the error. For instance, the cross-classification of highly comorbid diseases is a poor way of fixing the problem, because it does not tackle the source of error either in the classification system or in the classifying method.

Causal classifications, on the other hand, seem to appeal to a metaphysical notion of accuracy, which is problematic because we have no access to true taxonomies, even assuming there are such things. The proponents of causal classifications aim to reflect the 'real' interacting biological, psychological and environmental causal factors. According to psychiatrist Kendler (2012a), for instance, the causal approach allows one to define psychiatric disorders more "realistically" and "presents a much more realistic picture of what an etiologically based psychiatric nosology would really look like" (Kendler, 2012b, pp. 17-18).[8] Similarly, according to the proponents of the Research Domain Criteria (RDoC), a research project that aims to study the causes of neuropsychological constructs, "the pre-eminent role of diagnosis in medicine is to determine the exact nature of a patient's disease in order to administer the optimal treatment" (Cuthbert & Insel, 2013, p. 2).

The principle behind these claims is that a classification system should be evaluated according to how close it gets to the Platonic ideal of 'splitting nature at its joints' (*Phaedrus* 265c-266b). In the case of psychiatric taxonomy, this means assuming that disorders have theory-independent features that distinguish them from one another, and should be classified according to these differences. In philosophy of measurement, this principle is reflected in the metaphysical concept of error as distance from the true value of a quantity, which measurement outcomes are meant to approximate (Swoyer, 1987). Setting aside the metaphysical debates, the concept of true value is not useful for understanding the assessment of measurement in scientific practice, because scientists have no epistemic access to 'true' values and hence cannot possibly evaluate the accuracy of their measurements against them. Therefore, even assuming the existence of theory-independent distinctions between

[8] In other works, Kendler suggests that the revisions of the modern symptom-based classifications can generate a progressive and cumulative process, in which each version improves over the previous one. The condition for this to happen is that revisions are evidence-based improvements on previous versions (Kendler, 2012a, 2009). This idea draws on the same philosophical insights as my article. However, Kendler is not clear about what is the criterion for judging whether the changes are indeed improvements. In my framework, this criterion is given by the coherence between the abstract classification system and the concrete classifying method, which are simultaneously put to test by comparing different, but nevertheless related classificatory perspectives on mental disorders.

psychiatric disorders, scientists have no epistemic access to these junctions that is independent of any way to identify and classify these disorders.

Taking a measurement perspective, I suggest moving from operational and metaphysical notions of accuracy to an epistemic one. The notion of accuracy that emerges from measurement assessment practice has two main advantages. First, it is a metaphysically neutral notion of accuracy, which does not rely on having access to an independent reality. Instead, according to model-based accounts, accuracy is the closeness of agreement among multiple measurements of a certain quantity, given their respective models. On this view, accuracy can be thought of as the probability of error that is associated with the measurement procedure on the bases of a model of how the procedure works. Second, it comes with a strategy for evaluating and improving accuracy in measurement practice: scientists evaluate and improve the epistemic accuracy of their measurements by comparing outcomes to each other. Under this conception, reliability is one of the components of accuracy, which arises from the evaluation of uncontrolled variations of the indications over repeated measurements. Validity is another component or accuracy, which is related to sources of uncertainty like systematic errors and vague measurand definitions among others (Tal, 2020).

In the context of classification, epistemic accuracy is a notion that relates the classification system and the classifying method, and depends on the fit between the two. The accuracy of a classification is the probability of error that is associated to the classifying method based on a model of how the method ideally works, that is, based on a representation of how the classifier assigns individuals to the correct classes in the relevant classification system.

4.1 An assumption of local nomic coherence

To evaluate a classification according to this notion of accuracy, we need to compare how different classifying methods assign individuals to classes based on their representations. This can be done by comparing field specific perspectives on mental disorders. Recall that, under the model-based account, the comparison of different measurements is epistemically meaningful only if they are represented in terms of the same quantity of interest. In other words, to evaluate the epistemic accuracy of a classification, scientists can compare different classifying methods based on the assumption that they work under compatible classification systems. Under this assumption, the classifying methods are deemed accurate if they agree with each other within their uncertainty intervals: the convergence is taken as a sign that the methods do not depart from the ideal model more than what can be expected and possibly corrected for. In contrast, discrepancies can be interpreted as signs of errors and prompt investigation of their sources, and of possible ways to correct them. Accordingly, improving the agreement amounts to improving the accuracy of the classification.

In the case of psychopathology, this seems to imply that the comparison of field-specific perspectives on psychiatric disorders presupposes that mental disorders are represented in compatible terms across different disciplines, which is seldom the

case given that field-specific perspectives are often found to cross-cut one another. Recent philosophical works on the representation of mental phenomena, in fact, have shown that the integration of disciplinary perspectives face problems due to the lack of alignment between these field specific models (Stinson, 2016; Sullivan, 2008; Marchionni manuscript). If we could agree upon the classification system, then it would be easy to compare alternative classifying methods. Scientists could evaluate the accuracy of classifying methods in terms of their probability of providing mutually consistent diagnoses based on their respective representations. For instance, based on their models, two classifying methods could be deemed accurate if they provide mutually consistent diagnoses in 90% of the cases. Discrepancies that go beyond the expected uncertainty intervals, on the other hand, could be interpreted as signs of error and prompt investigation of their sources. The situation in psychiatry, however, is different, because there is extreme uncertainty about the classification system per se: there are multiple classifying methods *and* multiple classification systems. Different disciplinary perspectives have their own ways of splitting and lumping mental phenomena according to different organizing principles, as well as distinctive methods for assigning individuals to classes.

According to Tal (2019, p. 873), however, measurement practice shows that this assumption comes down to a "weak requirement for nomic coherence": two instruments can be seen as measuring the same quantity when the quantity parameter enters into approximately the same nomic relations with other theoretical parameters in the respective models. In other words, there is no need of a strong identity between quantity parameters before explorative comparisons can be made – and it could not be otherwise, given that the definition of the quantity is thought to co-evolve with its measurement. Indeed, Tal (2019) highlights that a very productive strategy employed by scientists involved in a cutting-edge measurement project is to compare measurements from different domains by dogmatically assuming that they measure the same quantity. Tal warns, however, that this dogmatic assumption "should not be confused with empirical knowledge, because novel measurements may lead to the discovery of new laws and to the postulation of general quantities that are different from those initially supposed" (Tal, 2019, pp. 874-875).

In the case of psychiatry, this view encourages the comparison of findings from different perspectives even without a previous effort to provide a unified understanding of the underlying causes of mental disorders. This exploration is epistemically justified by assuming, provisionally, that different field-specific perspectives bring about evidence of the same psychopathologies. In this view, the different ways in which psychiatric disorders are modelled and identified in various domains have the potential to reveal previously undetected errors and can strengthen the accuracy of the classification when convergences are obtained. This allows explorative comparisons of not fully consistent or even conflicting perspectives and emphasizes the epistemic potential of their differences.

The possibility of making explorative comparisons, however, does not dismiss the problem of nomic coherence: even if we proceed with explorative comparisons, we still need, at some point, to show the nomic coherence of the classifications. What would be the criteria for judging the nomic coherence of different, discipline-specific classifications of mental phenomena is not a straightforward matter. Presumably, it

would require the classifications to be compatible at some non-trivial level of generality, but the psychological, biological, and environmental perspectives on mental phenomena intersect in various ways, and this does not allow us to come up with a mapping between different perspectives. Whether and how the nomic coherence of alternative classifications can be evaluated in the context of psychiatric taxonomy remains an open question that cannot be solved here. However, while this continues to be a problem for comparing classification systems in their entirety, it might be possible to address nomic coherence at a local level. In other words, it might be feasible to evaluate the nomic coherence of specific mental phenomena as represented from different field-specific perspectives, even without comparing the entire classification systems where they belong. For instance, psychological and biological perspectives might approximately agree on the individuation of specific mental disorders, while continuing overall to split and lump mental phenomena in non-aligning or conflicting ways. Certain mental phenomena are characterized in similar ways across disciplines, for instance as being related to certain symptoms, as responding to certain kinds of treatments, as having certain epidemiological characteristics, etc., and this might allow for meaningful comparisons among disciplinary perspectives.

Based on the assumption of local nomic coherence, comparisons of different perspectives on specific mental phenomena might lead to accuracy improvements if they allow us to identify and correct errors. The next section provides an example of interdisciplinary research efforts into the aetiology of psychopathology and asks whether the insights from measurement assessment practice can help shed light on the epistemic bases of successful research in this field.

5 Improving accuracy via local comparisons: The epigenetics of gene-environmental interactions

The comparison of different disciplinary perspectives on specific mental disorders can be seen as providing opportunities for detecting errors and improving epistemic accuracy based on local representations of the interaction between causal factors. As an example, consider the recent studies about the epigenetic basis of gene-environment (G×E) interactions in the development of stress-related disorders (major depressive disorders (MDD), anxiety disorders, and post-traumatic stress disorder (PTSD)).

Both the environmental and the genetic perspectives have unaccounted sources of error in their identification of stress-related disorders. Epidemiological findings highlight the effect of the environment on the risk for psychiatric diseases, and especially the impact of early life trauma on stress-related disorders, but also acknowledge a considerable individual variability of the outcomes following exposure to traumatic experiences. From the genetic perspective, while there is a significant genetic contribution to the development of other psychopathologies such as schizophrenia and bipolar disorder, strong genetic effects have not been observed for stress-related disorders, as indicated by the lack of consistent and replicated findings in GWAS (Otte et al., 2016). These findings can be interpreted as signs of error, highlighted by the unaccounted variability in the two domains: there are unexpected

discrepancies between ideal model and actual outcomes in both disciplines. From the perspective of environmental risk factors, the high variability of the outcomes following exposure to trauma goes beyond what can be expected based on the field-specific representation. This means that two classifiers working within the environmental perspective would give outcomes that diverge more than the expected uncertainty intervals. Similarly, the lack of consistent replicated results in the genetic of stress-related disorders does not find an explanation in the field-specific representation. Within each field, the variability remains unaccounted for. Epidemiological studies comparing the genetic and environmental perspectives on these disorders, however, also highlight a combined contribution of genetic and environmental risk factors. Investigating their interaction, therefore, can help to throw light on previously unaccounted for variability in the two domains.

Recent studies investigating the molecular mechanisms underlying G×E interactions show that they might involve epigenetic regulation (Klengel & Binder, 2015; Klengel et al., 2013). This work is novel not because it brings together environmental and genetic factors, as has been done before in psychology and psychiatry, but because it provides an explanation of *how* they interact at the epigenetic level. In particular, these studies provide evidence that early life trauma can lead to lifelong molecular changes in the form of epigenetic modifications capable of shaping the response to environmental stress in adult life, thereby affecting the vulnerability to stress-related psychiatric diseases. This molecular mechanism of gene-environment interplay can explain the conditions under which traumatic experience in early life increases the risk for stress-related psychiatric disorders in adults.

The results of these studies can help detecting and correcting for errors in the genetic and environmental perspectives, leading to mutual refinements and better fit with the empirical findings. The epigenetic mechanism of G×E interplay can explain part of the heterogeneity in the responses to environmental adversity, because the long-term epigenetic changes depend on the characteristics of the stressor as well as on the interplay with genetic predisposition. On the other hand, the epigenetic mechanisms could be relevant for explaining the missing heritability observed in stress disorders (Klengel & Binder, 2015). The heterogeneity in genetic studies can be explained, at least in part, by the observation that relevant genetic variants increase the risk of developing stress-related psychiatric disorders only in the case of exposure to stressors or other adverse environmental circumstances (Otte et al., 2016).

In this example, the comparison of different perspectives not only leads to new attempts to explain the interactions between factors, but also provides new insights on how to correct for errors by refining the representations in the respective fields. The epigenetic findings, in fact, account for the errors in the environmental and genetic perspectives and offer the possibly to correct them, and thereby to improve the epistemic accuracy of both field-specific perspectives on mental disorders. If future epidemiological studies will show a better agreement once the representations have been corrected for, then this could be taken as a sign that epistemic accuracy has improved. This is done on the basis of 'local' representations of the interactions between genetic and environmental factors, and do not rely on a comprehensive, integrated model of the aetiology of mental disorders.

This kind of inter-field investigation, moreover, can provide insights for how to demarcate mental disorders, and can thereby contribute to improving the accuracy of their classification. The authors use symptom-based categories to select the population under study, but their findings confirm some aspects of this classification while suggesting, at the same time, that it might not be the best way to classify these disorders. On the one hand, the epigenetic findings ground the relevance of both genetic and environmental risk factors of stress-related disorders (because unaccounted variability in the two domains is explained away) and especially of their combined contribution. In this way, they mitigate doubts about the biological (genetic, at least) relevance of these symptom-based categories. On the other hand, however, the authors observe that the epigenetic interaction between genetic predisposition and early-life traumatic experience predisposes to both PTSD and MDD. This reveals that this interaction is likely to be relevant for stress sensitivity in general, crossing current diagnostic borders (Klengel et al., 2013). In other words, these disorders might be better classified as one single category. The implications of the study, therefore, go beyond the symptom-based classes used to begin with, and appear to be relatively independent of the opposition between causal and symptom-based approaches. Although a single study cannot provide definitive answers to these classificatory issues, its findings contribute to making progress possible.

The choice of using the general category of stress-related disorders rather than the more fine-grade distinctions found in the current classifications can be interpreted as a way to address the requirement of nomic coherence: the choice might be motivated by the need to find the level of generality at which the involved disciplines find their intersection point. In this view, the authors compare different disciplinary perspectives on a specific set of disorders, based on the assumption that these disorders are represented in similar ways across domains. This assumption does not imply more general commitments about the coherence among discipline-specific ways of demarcating between other mental phenomena.

In sum, this example illustrates how the epistemic accuracy of psychiatric classifications can be evaluated and improved by local efforts to compare findings from different perspectives without appealing to a unified representation of the aetiology of mental disorders. This strategy is applicable when there are different but nevertheless related classificatory perspectives on specific disorders, independently of the kind of discipline involved. What is needed for improving the classification is that different perspectives can be compared to each other, thereby allowing for mutual refinements of the classification systems and the classifying methods. The proponents of causal classifications recognize this as a successful strategy, but lament that the studies investigating the relations between multiple causal factors are relatively rare as compared to those focused on one single causal factor. Studying the interactions between factors can help identify local circumstances where it is fruitful to compare different disciplinary perspectives for improving the classification of mental disorders. This idea is in line with recent philosophical works on the multiple disciplinary perspectives on mental disorders, which argue that successful integration comes from uncovering connections between partially overlapping representations, rather than from unifying mechanisms (Stinson, 2016). My account supports this view and provides a systematic framework for understanding how the comparison

of findings from different disciplines can give rise to circumscribed but successful integration without relying on a unifying representation of mental phenomena.

6 Conclusions

The measurement framework provided in this paper suggests that a classification can be evaluated in terms of epistemic accuracy, that is, based on the mutual refinements of classifying method and classification system. To evaluate a classification according to this notion of accuracy, scientists can compare different classifying methods based on the assumption that they work under compatible classification systems. In this perspective, the coherence between classifying method and classification system is an alternative to both the appeal to an independent truth and to aiming only at the consistency of results over repetition.

The evaluation of psychiatric classification involves deep epistemic uncertainty, because both the classifying method and the classification system are at the centre of heated debates. As a consequence, it is not clear how scientists could make meaningful comparisons across disciplinary perspectives that have distinct methods for assigning individuals to classes and different classification systems that split and lump mental disorders in non-aligning ways. While this remains a problem for comparing classifications in their entirety, this paper argued that scientists are able to improve the classification of mental disorders by means of 'local' comparisons among perspectives on specific mental phenomena. Instead of relying on the alignment of different perspectives to provide a unified representation, the local comparison across fields explores the implications of the convergences and discrepancies between them. This allows to simultaneously test the classification systems and the classifying methods, which improve by mutually influencing each other.

On this account, the successful improvement of epistemic accuracy is relatively independent of whether mental phenomena are initially represented in terms of their symptomatology or of their aetiology. This is an effect of the requirement of local nomic coherence: since scientists compare phenomena that are represented in compatible ways across disciplines, the 'nomenclature' for the comparison is subordinate to how the phenomena are represented in the relevant fields. For instance, in the example discussed above, scientists compare genetic and environmental perspectives on stress-related disorders: they rely on the symptom-based categories of MDD, anxiety disorders, and PTSD for selecting the population under study, but they investigate the causal factors of the mental phenomenon of stress-sensitivity in general. Indeed, their results provide insights into how to classify mental phenomena that go beyond the debate between symptom-based and causal classification: on the one hand, the epigenetic findings mitigate doubts about the genetic relevance of the employed symptom-based categories; on the other hand, they suggest that these categories might be better treated as one single mental phenomenon. As a consequence, the measurement perspective provided here suggests that the recent psychiatric debates have been overly concerned with the dichotomy of symptom-based

versus causal approaches, which, despite being central to explanation and understanding, is not essential to improving the accuracy of psychiatric classification.[9]

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest/competing interests The author declares that there are no conflicts of interest.

References

Alexandrova, A. (2016). Is well-being measurable after all? *Public Health Ethics*, phw015. https://doi.org/10.1093/phe/phw015

APA. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). American Psychiatric Association.

Basso, A. (2017). The appeal to robustness in measurement practice. *Studies in History and Philosophy of Science Part A, The Making of Measurement, 65–66*, 57–66. https://doi.org/10.1016/j.shpsa.2017.02.001

Blashfield, R. K., Keeley, J. W., Flanagan, E. H., & Miles, S. R. (2014). The cycle of classification: DSM-I through DSM-5. *Annual Review of Clinical Psychology, 10*, 25–51. https://doi.org/10.1146/annurev-clinpsy-032813-153639

Bolton, D. (2012). Classification and causal mechanisms: A deflationary approach to the classification problem. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 6–11). Oxford University Press. https://doi.org/10.1093/med/9780199642205.003.0002

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*(9), 1089–1108. https://doi.org/10.1002/jclp.20503

Boumans, M. (2015). *Science outside the laboratory: Measurement in field science and economics*. Oxford University Press.

Bridgman, P. W. (1927). *The logic of modern physics*. Macmillan.

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology, 26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Byerly, H. C., & Lazara, V. A. (1973). Realist foundations of measurement. *Philosophy of Science, 40*(1), 10–28. https://doi.org/10.1086/288493

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

[9] This is not equivalent to arguing that classificatory issues are unimportant in the discovery of causal mechanisms (cf. Bolton, 2012).

Chang, H. (2004). *Inventing temperature: Measurement and scientific Progress*. Oxford University Press.
Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's research domain criteria (RDoC). *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 18*(2), 72–145. https://doi.org/10.1177/1529100617727266
Compton, W. M., & Guze, S. B. (1995). The neo-Kraepelinian revolution in psychiatric diagnosis. *European Archives of Psychiatry and Clinical Neuroscience, 245*(4–5), 196–201. https://doi.org/10.1007/bf02191797
Cooper, R. (2018). *Diagnosing the diagnostic and statistical manual of mental disorders* (5th ed.). Routledge. https://doi.org/10.4324/9780429473678
Cuthbert, B. N. (2015). Research domain criteria: Toward future psychiatric Nosologies. *Dialogues in Clinical Neuroscience, 17*(1), 89–97.
Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine, 11,* 126. https://doi.org/10.1186/1741-7015-11-126
Demazeux, S., & Singy, P. (Eds.). (2015). *The DSM-5 in perspective: Philosophical reflections on the psychiatric babel. History, philosophy and theory of the life sciences*. Springer. https://doi.org/10.1007/978-94-017-9765-8
Ellis, B. (1966). *Basic concepts of measurement*. Cambridge University Press.
Follette, W. C., & Houts, A. C. (1996). Models of scientific Progress and the role of theory in taxonomy development: A case study of the DSM. *Journal of Consulting and Clinical Psychology, 64*(6), 1120–1132. https://doi.org/10.1037//0022-006x.64.6.1120
van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.
Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese, 175*(2), 123–149. https://doi.org/10.1007/s11229-009-9466-3
Fulford, KWM, and Norman Sartorius. 2009. "A secret history of ICD and the hidden future of DSM." In *psychiatry as cognitive neuroscience: Philosophical perspectives*, edited by Matthew Broome and Lisa Bortolotti. Oxford University Press.
Guyatt, G. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA, 268*(17), 2420. https://doi.org/10.1001/jama.1992.03490170092032
Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. Wiley.
Hyman, S. E. (2007). Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience, 8*(9), 725–732. https://doi.org/10.1038/nrn2218
Joint Committee for Guides in Metrology (JCGM). (2012). *International vocabulary of Metrology - basic and general concepts and associated terms (VIM)* (3rd ed.). JCGM.
Kendler, K. S. (2005). Toward a philosophical structure for psychiatry. *American Journal of Psychiatry, 162*(3), 433–440. https://doi.org/10.1176/appi.ajp.162.3.433
Kendler, K. S. (2009). An historical framework for psychiatric nosology. *Psychological Medicine, 39*(12), 1935–1941. https://doi.org/10.1017/S0033291709005753
Kendler, K. S. (2012a). Epistemic iteration as a historical model for psychiatric nosology: Promises and limitations. In K. S. Kendler & P. Josef (Eds.), *Philosophical issues in psychiatry II: Nosology*. Oxford University Press.
Kendler, K. S. (2012b). Levels of explanation in psychiatric and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry, 17*(1), 11–21. https://doi.org/10.1038/mp.2011.70
Kendler, K. S., & First, M. B. (2010). Alternative futures for the DSM revision process: Iteration v. paradigm shift. *The British Journal of Psychiatry: The Journal of Mental Science, 197*(4), 263–265. https://doi.org/10.1192/bjp.bp.109.076794
Kendler, K. S., & Parnas, J. (2012). *Philosophical issues in psychiatry II: Nosology*. Oxford University Press.
Kendler, K. S., & Parnas, J. (2017). *Philosophical issues in psychiatry IV: Psychiatric nosology*.
Klengel, T., & Binder, E. B. (2015). Epigenetics of stress-related psychiatric disorders and gene × environment interactions. *Neuron, 86*(6), 1343–1357. https://doi.org/10.1016/j.neuron.2015.05.036
Klengel, T., Mehta, D., Anacker, C., Rex-Haffner, M., Pruessner, J. C., Pariante, C. M., & Thaddeus W. W. Pace, et al. (2013). Allele-specific FKBP5 DNA demethylation mediates gene–childhood trauma interactions. *Nature Neuroscience, 16*(1), 33–41. https://doi.org/10.1038/nn.3275

Kupfer, D. J., & Regier, D. A. (2011). Neuroscience, clinical evidence, and the future of psychiatric classification in DSM-5. *The American Journal of Psychiatry, 168*(7), 672–674. https://doi.org/10.1176/appi.ajp.2011.11020219

Lalumera, E. (2016). Saving the DSM-5? Descriptive conceptions and theoretical concepts of mental disorders. MEDICINA & STORIA, 109-128. *Medicina E Storia, 9–10*, 109–129.

Marchionni, Caterina. (n.d.) Unpublished manuscript. "Challenging the Mechanistic View of Psychiatry."

Mari, L. (2005). Models of the measurement process. In P. Sydenman & R. Thorn (Eds.), *Handbook of measuring system design* (2:Ch. 104). Wiley. https://doi.org/10.1002/0471497398.mm066

Mari, L., & Wilson, M. (2014). An introduction to the Rasch measurement approach for Metrologists. *Measurement, 51*, 315–327. https://doi.org/10.1016/j.measurement.2014.02.014

McClimans, L. (2017). *Measurement in medicine: Philosophical essays on assessment and evaluation*. Rowman & Littlefield International.

McClimans, L., Browne, J., & Cano, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Studies in History and Philosophy of Science Part A, The Making of Measurement, 65–66*(October), 67–73. https://doi.org/10.1016/j.shpsa.2017.06.004

Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person Back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective, 2*(October), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Murphy, D. (2006). *Psychiatry in the scientific image. Philosophical psychopathology*. MIT Press.

Otte, C., Gold, S. M., Penninx, B. W., Pariante, C. M., Etkin, A., Fava, M., Mohr, D. C., & Schatzberg, A. F. (2016). Major Depressive Disorder. *Nature Reviews. Disease Primers, 2*, 16065. https://doi.org/10.1038/nrdp.2016.65

Sackett, D. L. (1997). Evidence-based medicine. *Seminars in Perinatology, Fatal and Neonatal Hematology for the 21st Century, 21*(1), 3–5. https://doi.org/10.1016/S0146-0005(97)80013-4

Sackett, D. L. (Ed.). (2000). *Evidence-based medicine: How to practice and teach EBM. 2nd ed., reprinted*. Churchill Livingstone.

Staley, K. W. (2020). Securing the empirical value of measurement results. *The British Journal for the Philosophy of Science, 71*(1), 87–113. https://doi.org/10.1093/bjps/axx036

Stengel, E. (1959). Classification of mental disorders. *Bulletin of the World Health Organization, 21*, 601–663.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Stinson, C. (2016). Mechanisms in psychology: Ripping nature at its seams. *Synthese, 193*(5). https://doi.org/10.1007/s11229-015-0871-5

Sullivan, J. A. (2008). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the Unity of neuroscience. *Synthese, 167*(3), 511–539. https://doi.org/10.1007/s11229-008-9389-4

Swoyer, C. (1987). The metaphysics of measurement. In J. Forge (Ed.), *Measurement, realism and objectivity: Essays on measurement in the social and physical sciences* (pp. 235–90). Australasian Studies in History and Philosophy of Science. Springer Netherlands. https://doi.org/10.1007/978-94-009-3919-6_8

Tabb, K. (2017). Philosophy of psychiatry after diagnostic kinds. *Synthese, 196*. https://doi.org/10.1007/s11229-017-1659-6

Tal, E. (2011). How accurate is the standard second? *Philosophy of Science, 78*(5), 1082–1096. https://doi.org/10.1086/662268

Tal, E. (2016). Making time: A study in the epistemology of measurement. *The British Journal for the Philosophy of Science, 67*(1), 297–335. https://doi.org/10.1093/bjps/axu037

Tal, E. (2017). A model-based epistemology of measurement. In N. Mößner & A. Nordmann (Eds.), *Reasoning in measurement* (pp. 233–253). Routledge.

Tal, E. (2019). Individuating quantities. *Philosophical Studies, 176*(4), 853–878. https://doi.org/10.1007/s11098-018-1216-2

Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Tsou, J. Y. (2015). DSM-5 and psychiatry's second revolution: Descriptive vs. theoretical approaches to psychiatric classification. In S. Demazeux & P. Singy (Eds.), *The DSM-5 in perspective: Philosophical reflections on the psychiatric babel* (pp. 43–62). History, Philosophy and Theory of the Life Sciences. Springer Netherlands. https://doi.org/10.1007/978-94-017-9765-8_3

WHO. (1967). *International statistical classification of diseases, injuries, and causes of death, 8th revision*. World Health Organization.

WHO. (1974). *Glossary of mental Disor- Ders and guide to their classification*. World Health Organization.

Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement, 46*(9), 3766–3774. https://doi.org/10.1016/j.measurement.2013.04.005

Wilson, M. (1993). DSM-III and the transformation of American psychiatry: A history. *The American Journal of Psychiatry, 150*(April), 399–410. https://doi.org/10.1176/ajp.150.3.399

Zachar, P., Stoyanov, O. S., Aragona, M., & Jablensky, A. (Eds.). (2014). *Alternative perspectives on psychiatric validation: DSM, ICD, RDoC, and beyond. International perspectives in philosophy and psychiatry*. Oxford University Press.