Measurement 225 (2024) 113993

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

The role of ethical and social values in psychosocial measurement

Sebastian Rodriguez Duque [a,1,*], Eran Tal [a,2], Skye Pamela Barbic [b,c,d]

[a] Department of Philosophy, McGill University, 855 Sherbrooke Street West, Montreal, Quebec H3A 2T7, Canada
[b] Occupational Science and Occupational Therapy, Faculty of Medicine, University of British Columbia, T325-2211 Westbrook Mall, Vancouver, B.C. V6T 2B5, Canada
[c] Centre for Advancing Health Outcomes, 570-1081 Burrard Street, St. Paul's Hospital, Vancouver, B.C. V6Z 1Y6, Canada
[d] Foundry British Columbia, 1045 Howe Street, Vancouver, B.C. V6Z 2A9, Canada

ARTICLE INFO

Keywords:
Psychosocial measurement
Philosophy of science
Values in science
Mental health
Psychometrics
Integrated youth services
Evaluation

ABSTRACT

In the natural sciences, measurement is taken as a reliable source of knowledge because it requires a special kind of rigour. This is usually understood as an epistemic kind of rigour. We argue that in psychometrics, the science of measuring mental traits, attitudes, and experiences, the quality of knowledge supplied by measurement procedures must be established through ethical as well as epistemic justification. We reject views that restrict the role of ethics in measurement to guarding against potential harmful consequences and show that ethical and social value judgments are intrinsic to the design, validation, interpretation, and use of psychometric tools. We propose a five-step procedure called 'ethical iterations' that allows researchers, decision makers, and other interested parties to ensure that measurement practice is aligned with their aims and values. We substantiate our claims with evidence from our work with Foundry, a youth health organization in British Columbia, Canada.

1. Introduction

Psychometrics, the science of measuring mental traits, attitudes, and experiences, guides the design and use of measuring instruments in a variety of fields, including psychology, education, healthcare, and management. Psychometric measures, such as questionnaires and tests, are meant to evaluate an attribute ('construct') based on recorded human behaviour, such as responses to a questionnaire. In the field of mental health measurement, constructs are often modelled after categories in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM),[3] such as anxiety and depression. The results of measuring such constructs inform high-stakes decisions concerning treatment, diagnosis, insurance coverage, and eligibility for benefits, among others. With notable exceptions we discuss below, the psychometric and philosophical literature analyzes the quality of psychometric measures almost exclusively in epistemic terms. That is, based on their validity, reliability, sensitivity, and specificity. We argue that epistemic evaluation is insufficient for the coordination of psychosocial[4] measurands with the instruments that purport to measure them. Ethical considerations are also required.

Ethical considerations are identified as important in some key strands of psychometric theory, most notably the theoretical tradition informing the *Standards for Educational and Psychological Testing* ([2]; e.g. Kane [3]; Messick [4]; Shepard [5,6]). The *Standards* frame ethical considerations as issues of *fairness* related to the management of bias and the mitigation of negative consequences of test administration and use. We agree that the emphasis on fairness developed by the *Standards* is

* Corresponding author.
E-mail addresses: sebastian.rodriguezduque@mail.mcgill.ca (S. Rodriguez Duque), eran.tal@mcgill.ca (E. Tal), skye.barbic@ubc.ca (S. Pamela Barbic).

[1] Leacock Building, Room 414.
[2] Leacock Building, Room 933.
[3] Now in its fifth edition, the latest version at the time of writing is the DSM-5-TR [1].
[4] Our example in this paper centres on considerations arising in the field of mental health. Besides clinical outcome assessments, such as patient reported outcomes (PROMs), there are other measures that are relevant to this field, such as summary measures of population mental health that rely on system-wide indicators. These include, for example, the number of emergency visits, number of hospitalizations, and number of overdose deaths for a given population. We take the term 'psychosocial' and our argument concerning the role of ethical and social values to apply to these summary measures as well. However, this paper will focus on measurements of properties of individuals rather than of populations.

https://doi.org/10.1016/j.measurement.2023.113993
Received 14 February 2023; Received in revised form 21 August 2023; Accepted 4 December 2023
Available online 6 December 2023

important. However, we argue that attention to the role of values needs to go beyond managing them as potential sources of bias, or to guard against harmful consequences. We emphasize the intrinsic role of ethical and non-epistemic[5] value considerations for the definition and interpretation of psychosocial measurands. Proper theorizing and operationalization of psychometric constructs must include an ethical evaluation, and this ethical evaluation requires the explicit consideration of non-epistemic values. Ethical considerations play a role, not only in ensuring that measurement is fair and avoids undesirable social consequences, but also in shaping the content of the construct, aligning measurement practice with the values of interested parties, facilitating a productive and respectful conversation with respondents, and promoting individual and social goods, such as wellbeing and equity. These broader roles for ethical values are consistent with holistic views of psychometric validity, such as those developed by Messick and Kane, but have been difficult to translate into methodology. Consequently, these broader roles are rarely acknowledged in practical guides like the *Standards* or made explicit as part of the development and use of psychometric measures. This article will outline a procedure for conducting explicit, comprehensive, and evidence-based deliberations on the ethical and social values that inform a given measurement practice.

Our methodology combines qualitative empirical work with insights drawn from the psychometric literature and the philosophy of measurement. First, we discuss the role of values at large in securing the coordination of measurement instruments with their measurands. Then we discuss the specific case of mental health measurement in a clinical setting. The case study is based on our interdisciplinary collaboration, which involved philosophers, a health outcomes researcher, and service providers who serve youth in British Columbia. We identify and analyze three gaps between psychometric theory and clinical practice that became evident from this collaboration. These gaps reveal the central role of non-epistemic values in psychosocial measurement practice and point to a need for a framework that makes their role explicit and productive. A key element of this framework is a process we call 'ethical iterations'. During ethical iterations, the values informing a measurement practice, and the practice itself, are mutually refined in light of evidence. Our final section elaborates on this notion and points to its desired result: a measurement practice that is fit-for-purpose.

2. The authority of measurement

Measurement is widely viewed as a reliable source of knowledge. This reliability provides measurement results with epistemic authority, as demonstrated by their widespread use as evidence in scientific and practical settings. Yet the sources of reliability of measurement require clarification. As has been recently noted by Luca Mari et al. [9], "… measurement is regarded as integral to science and society on the basis of its epistemic authority, and so the question remains of what, exactly, justifies claims to such authority…?" [9], p. 5. We set out to help answer this question. In particular, while an authoritative measurement process in the physical sciences often requires non-epistemic value judgments (e.g., in the management of inductive risk), non-epistemic values[6] play an even stronger role in the definition of psychosocial measurands and the practices of measuring them.

Measurement is viewed as authoritative because it is understood to require a special kind of rigour. This is usually understood as an epistemic kind of rigour. At the heart of this rigour is the goal of the successful coordination of quantity concepts with the instruments used to measure them. The coordination of a quantity concept with its methods of measurement is a gradual and iterative process. This is a key lesson from recent historical and philosophical studies of measurement, most notably by Hasok Chang [10] and Bas van Fraassen [11]. These authors focus largely on measurement in the natural sciences and emphasize the epistemic aspects of iterations. As part of the coordination of a quantity concept like time or temperature with its methods of measurement, theories of the quantity and instruments for measuring that quantity are mutually refined with the aim of achieving a coherent fit. For example, the question 'what counts as the correct temperature of an object?' was historically answered iteratively and simultaneously with the question 'what counts as an accurate thermometer?' With each iteration, some prior traditions and beliefs concerning temperature and thermometry were retained, while others were revised in order to resolve inconsistencies and reduce measurement error. But in order to escape a possible vicious circularity, these processes and procedures, as a practice, need to be shaped and evaluated in light of the right epistemic values. Chang discusses epistemic values he calls 'progress' and 'respect', which enable the enrichment and self-correction of an iterative process [10], p. 228. Therefore, even in physical measurement, values play a decisive role in the rigour of measurement practice, which grounds its authority.

The framework of epistemic iteration is a successful account of coordination for physical measurement. Ethical, social or political values do not play a role in the theoretical description of temperature or time in the same way they do for psychosocial measurands associated with wellbeing or mental health. And the ethical stakes of solving the problem of coordination one way or another are lower. For example, there is little at stake from an ethical or social value perspective as to whether temperature is theorized as a kinetic or thermodynamic property,[7] and, crucially, it would be hard to make the case that such theories are constitutively shaped by those values. Likewise, although many high-stakes decisions depend on the ability to measure temperature accurately, there is little at stake from an ethical or social perspective as to which type of thermometer turns out to be best for the job – say, a bulb, gas, resistance, or infrared thermometer. This is not to say that the comfort of a patient may not be better served depending on the kind thermometer used to measure their fever.[8] This is to say that the coordination process of temperature and thermometers will be largely unaffected by how the patient experiences that interaction.

In the case of a measurand like depression, the instrument, the circumstances, and the character of the interaction itself affect the measurement outcome and its interpretation and interpretability for a given person. The purpose and context of measurement affect the meanings attributable to an instrument indication (e.g. the score of a questionnaire) by test administrators and respondents; the inferences that may be drawn from the meaning that is attributed to instrument indications

[5] While the distinction between epistemic and non-epistemic values may be controversial [7] we maintain it here to emphasize the intrinsic social, ethical, and political dimensions of value judgments involved in psychometric theory and practice. Nothing in our argument hangs on whether the distinction between epistemic and non-epistemic values is defensible. For a discussion of the usefulness of this distinction see [8].

[6] Here we will use the term 'value' in its broadest axiological sense. For our purposes, values are normative guides and constraints on what matters in a given domain. We understand these as constraints not only because they are demanding, but because they bound the limits of action choices. In this discussion, nothing depends on whether something is finally valuable, intrinsically valuable, instrumentally valuable, or some combination of the former.

[7] This is not to deny that ethical and social consequences could follow even in the case of temperature. There may be interests related to, e.g., reputation, trust, safety, financial cost, and environmental impacts riding on which theories and instruments are deemed better than others. The ethical stakes in the resolution of a coordination problem should be judged on a case-by-case basis.

[8] We thank David Sherry for pressing this point: ethical dimensions are immediately relevant even in the use of some measurement instruments like thermometers, especially if what is being measured is an attribute of a human.

in that context; the actions that meaning will support; and whether or not instrument indications have any operative meaning at all.[9] As we discuss in the next section, when the properties being measured are human and social attributes, such as depression, reading ability, or quality of life, the attribute itself is defined and operationalized in light of non-epistemic values, and people typically have much to gain or lose depending on how the attribute is defined and operationalized. In these cases, ethical, social, and political considerations have legitimate roles in informing the process of coordination and in justifying the choice of theories and instruments used to conceptualize and measure the attribute. Epistemic and ethical considerations are both intrinsic to the definition and use of psychosocial measurands. Therefore, the authority of psychosocial measurement depends not only on epistemic, but also on ethical justification.

3. Making room for non-epistemic values in measurement

There are four main reasons why ethical and social values have a legitimate role in the coordination of psychosocial measurands with concrete measurement methods. First, measurement in the social and behavioural sciences often has normative ends, e.g., the improvement of human welfare. Second, the nature of constructs in the social and behavioural sciences is often such that ethical and social values are constitutive of their content and necessary for choosing among their possible operationalizations. Third, humans are often affected – positively or negatively – by the very act of being measured. Fourth, theoretical models of the measurement process tend to be significantly less detailed in the social and behavioural sciences than they are in the natural sciences. This leads to looser constraints on the right way to make progress on a coordination problem, and leaves room for ethical and social values to guide its solution. We will now briefly discuss each of these reasons in turn.

The first reason is straightforward. As one of the authors' undergraduate political science professors was fond of reminding his students: political scientists are interested in the phenomena of war and peace in part because they are interesting subjects of study, but more importantly because they want less of the first and more of the second. Often, philosophical and scientific interest in human and social phenomena is intrinsically normative; researchers and practitioners are working to ameliorate human lives and to relieve human suffering.

The second reason builds and expands on this first reason. The value-ladenness of measurement and classification in the social sciences has long been acknowledged by philosophers.[10] Cartwright and Runhardt [18] argue that the act of demarcating the content of a construct in the social sciences is value-dependent. For example, they claim that "civil war is not something that has definite boundaries nor, it seems, is there some one set of characteristics that all things we label as civil wars have in common" [18], p. 268. Following Otto Neurath, they call such concepts *Ballung* concepts, "…concepts that are characterized by family resemblance between individuals rather than by a definite property" [18], p. 268. The choice among the multiple meanings of such concepts is value-laden and enters each stage of the life of a measure, as we discuss below. The upshot of these reasons is that both the purposes of measurement and the content of the construct one aims to measure are constituted normatively in relation to ethical and social values.

Many of the human attributes being measured in psychology, healthcare, education, and economics are *Ballung* concepts. Some of them, such as quality of life, health-related quality of life, and happiness, wear their normativity on their sleeve: the concept of a good life is a central topic of investigation by moral theorists. Other measurands, such as anxiety or physical mobility, are normative in a more subtle sense. Scientists who define and operationalize such measurands need to choose which of their several meanings to take into account, and how to aggregate these meanings into a single number (or a handful of numbers). Different communities, organizations and people value different meanings of the same measurand and are variously affected by the consequences of including or excluding such meanings in a questionnaire or weighting them differently. As has been discussed by Hausman [19], p. 121, the effect of a broken finger on overall wellbeing may be very different for a philosopher than it would be for a professional violinist. Whether a measure of physical functioning queries the relevant information depends on what a given person, organization, or community values. The choice as to which items on a questionnaire are relevant and how to balance them is therefore intrinsically value-laden [20]. In the case of health-related measurement, examples of values that affect the content and wording of questionnaires are equity, inclusiveness, patient autonomy, the cost-efficiency of health services, and the transparency and accountability of healthcare providers, among others.

Scientists who demarcate the content of a concept of a human or social attribute and select how to operationalize it, thereby commit themselves – either explicitly or implicitly – to specific ethical and social values. Whether or not a measure is successful is determined in part by whether the ethical and social values that underlie its design are normatively justified, and by whether these values are a good fit for the purpose for which the measure will be used. Psychometricians in the field of educational assessment have long recognized this and have developed theories of validity that encompass the ethical dimensions and social consequences of a test. According to Messick [4,21] and Kane [3,12], validity is not a property of a test, but an evaluation of the collective evidence in favour of a specific use or interpretation of the instrument within a given context. The emphasis of this view on the ethical dimensions of measuring is very much aligned with our own. If health questionnaires are to serve as evidence for a wide array of decisions that promote social goods, such as improvements in healthcare access and quality, reduction of health disparities, and improved quality of life for patients, they must be based on value choices that promote these ends.

As we will discuss below, much of the ethical 'work' in defining and operationalizing healthcare constructs is currently done implicitly. Health outcome researchers are seldom explicit in their publications about the ethical and social values that guide their selection of content and wording for questionnaire items, and how these values fit the clinical purpose and target population.[11] One of our aims in this paper is to provide the conceptual groundwork for making such value choices explicit, and for assessing how well they fit with the intended use of the measure.

The third reason concerns how people may be affected by taking part in a measurement practice. Besides the intended use of a measuring instrument, there may be unintended consequences as well. This is especially the case when the entity is a person, and the instrument is a self-reported measure, such as a patient-reported outcome measure (PROM). To respond to a questionnaire, a person is required to interpret and reflect on a list of questions or statements. As Ian Hacking has argued, humans respond to attempts to classify them. This can lead to 'interactive kinds', that is, classes whose meaning and scope shift as subjects' behaviours, self-image, and social values change [23]. But even

[9] See e.g. Kane [3], [12] for his view of the argument-based approach to validity, which emphasizes the purpose and context of use as essential to ground the validity of inferences based on test interpretation. See also Larroulet Philippi [13] for his view that the notion of validity only makes sense in relation to a measurement purpose.

[10] Studies of the interdependence of measurement and social values include Hacking [14], Porter [15] and Gould [16]. For a recent discussion of the value-ladenness of psychometrics see Wijsen et al., [17]. For a helpful overview of the values in science literature see Elliott [7].

[11] See Alexandrova and Fabian's recent work [22], addressed further below, for a discussion on the different strategies available to researchers when managing evaluative judgments.

in the absence of such 'looping effects', questionnaires transmit beliefs and values to the persons being measured. The wording and choice of questions inform respondents about what service providers (or other experts) value, and about how experts view the content of the construct. These views may influence or clash with the views of respondents. Questions may be biased towards specific gender, age, race, ethnicity, or socio-economic groups, and thus be perceived as offensive or irrelevant by members of other groups. The absence of certain questions can also impact respondents: if a screening questionnaire for depression does not ask about social media use, a respondent who believes the long time they spend on social media aggravates their mental health may come to doubt themselves or distrust their healthcare provider. Privacy and data governance concerns are also central to the experience of answering health questionnaires. The respondent may rightly wonder who will own their data, who will be able to access their data, and for what purposes. Finally, the experience of being measured or classified can be harmful in that it could lead respondents to mis-conceptualize or reify their own mental traits ('I answered all questions with 'most of the time', therefore I must be irreparably mentally ill').

As the above points illustrate, the act of measuring often affects respondents, sometimes in subtle and unintended ways. As such, measurement can be understood as an intervention that affects respondents in two main ways. The first is the straightforward way of providing a measurement outcome that is interpreted by the people using the measure. For example, service providers and patients make decisions based on their interpretation of scores on health questionnaires. Patients may be affected for better or for worse by these interpretations and the resulting decisions. Truijens' et al.'s [24] recent work describes how the very act of measurement affects the way two respondents perceived their own symptoms as a result of their measurement experience. This mode of intervention occurs at the use stage of the measurement practice. The second mode occurs in the design stage of the practice. For example, designing an instrument around a DSM category, such as major depressive disorder, shapes and endorses a certain understanding of mental distress. This understanding then 'trickles down' to the organizations who use the measure and the patients who respond to it. In this way, instrument designers also intervene in the world when they shape the contours of psychosocial measurands. We will return to this point below when we discuss the harms of reification.

The combined upshot of the above considerations is that psychosocial measurement embeds non-epistemic values at many points: when constructs are defined, when instruments are designed and tested, when selecting which of several instruments to use, during administration, when interpreting scores, and when making decisions based on interpreted scores. It is not surprising that interested parties have a lot riding on the specific way a coordination problem is resolved in the behavioural and social sciences. The way the relevant mental trait or social category is theorized, the choice of method for measuring or classifying individuals under the trait or category, and the way the resulting data are interpreted are all value-laden and potentially high-stakes issues. Epistemic considerations alone cannot close this gap and determine the correct way to theorize and measure the construct. While epistemic considerations such as coherence, consistency, accuracy, predictive power, and explanatory power are all important desiderata in the development of measuring instruments, in the behavioural and social sciences they are insufficient to settle questions about whether and how well the instrument measures the intended construct.

Part of the reason for this insufficiency is the fourth reason listed above, that theoretical models of the measurement process are far less detailed in the behavioural and social sciences than they are in the physical sciences. In the physical sciences, the most accurate measuring instruments are modelled theoretically and statistically, taking into account a variety of interactions between the instrument, the object being measured, and the environment. Such models are often used to create an uncertainty budget that lists each of the known factors that influence the indications of an instrument, as well as an evaluation of the extent of bias caused by each factor and the uncertainty associated with the estimation of each bias [25,26]. Such level of detail is generally unavailable for the much more complex interaction between, say, a questionnaire, a person, and the circumstances under which the person responds to the questionnaire. Uncertainty budgets are cross-checked against each other, resulting in a tight web of constraints on what counts as an accurate instrument. These constraints make inconsistencies relatively easy to detect and allow scientists to formulate clear criteria for progress. By contrast, psychometric models are usually based on very loose theories of response behaviour that do not yet allow clear criteria of error and progress to emerge [27], p. 167. This means that progress in the design, use and interpretation of psychometric instruments can take multiple, divergent trajectories. The contingency of their evolution is more radical than that of many instruments in the physical sciences as described by Hasok Chang. This leaves room for non-epistemic values to influence the trajectory of coordination.

Despite the complexity of response behaviour, more detailed models are becoming available. One important example is the growing attention to the phenomenon known as response shift [28–30]. As defined by Sprangers and Schwartz [28], p. 1508, response shift, "… refers to a change in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the respondent's internal standards of measurement…; (b) a change in the respondent's values… or (c) a redefinition of a target construct." Techniques to investigate response shift, in particular appraisal models, shed light on the cognitive process of respondents' interpretation of a target construct in a questionnaire, and how such processes could change over time. Theories of response shift help make sense of observed changes in scoring where a person's health status has apparently deteriorated but their evaluation of their own quality of life has improved. Moreover, they help to distinguish between legitimate re-interpretations of the construct, and what may be mistaken interpretations of the construct.

One might therefore object that the undertheorized status of respondent behaviour may just reflect the state of the science in relation to psychosocial measurement in areas like psychiatry or psychology. Surely once finer-grained theories are developed, the need to appeal to non-epistemic criteria will be minimized. But even if a fine-grained theory existed that specified how different magnitudes of depression, say, give rise to different responses on a questionnaire, and how intervening factors mediate this process, epistemic factors would still be insufficient to decide which depression questionnaires are valid. This is because depression is a *Ballung* concept. Whether or not the hypothetical fine-grained theory concerns the right meaning of 'depression' depends on the specific characterization of depression and the purpose for which measurement is undertaken, and therefore also on the ethical and social values of the interested parties affected by the measurement.

In the remainder of the paper, we will proceed as follows. We will first provide a short survey of the literature on ethical values in psychometrics. Next, we will present a case study of measurement practice in a youth healthcare service to substantiate our claims. Then we will discuss how the four reasons identified above are exemplified in the case study. Finally, we will characterize ethical iterations as a structured way of making values explicit and integrating them into the coordination of a construct with a measurement practice.

4. Ethical dimensions of psychometrics

The centrality of values and the ethical dimensions of psychosocial measurement have been recognized in psychometrics since at least Samuel Messick's work [4,21,31], and have been a central component of the theory that informs the *Standards of Educational And Psychological Testing* [2]. As mentioned, the *Standards* focuses on issues of fairness in testing and the consequences of test use as part of the assessment of the validity of a measure. The work of authors like Kane [3,12] and Shepard [5,6] emphasizes that the validity of measures is relative to a purpose and context of use. Moreover, they also emphasize the role of values and

the social role of measurement practice as a whole, including the impact measurement has on those measured. In particular, Shepard [5] re-emphasizes Messick's [4] model of unitary validity, or the 'unified validity framework' [4]. In it, value implications and social consequences play a decisive role in a unified validity judgment about a given measure for a given purpose.[12] In this case, value implications map on to our previous discussion of the operationalization of *Ballung* concepts as being necessarily value dependent. 'Value implications' are the values embedded in the measure by virtue of the normative aims and theoretical commitments of researchers. Social consequences are the desirable or undesirable consequences of test use for a given purpose or context. In the case of education, for instance, whether a measure is valid would not only depend on whether it measures what it purports to measure. It would also have to be shown that no negative differential effects to the target population will ensue because of the introduction of the measure. The acknowledgement of the changing circumstances and various purposes for measures also leads these authors to stress the ongoing nature of validation.

Attention to the role of values in science has become a central concern in philosophy of science. They are, for example, important when selecting between theories that are underdetermined by evidence [33]. They are also exemplified by the role of non-epistemic values in the management of inductive risk [34]. Recently, the values embedded in psychometrics as a discipline have been discussed by Wijsen et al. [17]. This is an important task since, as they point out, "…psychometrics has traditionally been deeply invested in social and political projects, such as the eugenics movement, the introduction of military testing during the world wars, and the rise of a national education system" (See also, [35–38]). For example, in *The Mismeasure of Man*, Stephen Jay Gould [16] details the dubious social ends and ethical pitfalls associated with the development of the first 'intelligence' tests. We build on Wijsen et al.'s call to make explicit and critically reflect on the values embedded in psychometric practice, and the measurement of psychosocial attributes more broadly.

In the philosophical literature, important work is being done to acknowledge the role of ethical considerations alongside epistemic considerations in psychosocial measurement and the value-ladenness of concepts like well-being [22,30,39]. In particular, these authors are concerned with how to theorize the process of design and use of measures in a way that is attentive to the twin concerns of epistemic and ethical demands. McClimans [30,40] characterizes this process as a hermeneutic circle, which can be vicious or virtuous depending on the quality of epistemic dialogue engaged in by interested parties, and especially by virtue of empowering marginalized voices and promoting an inclusive exchange. This hermeneutic circle is similar to the iterative coordinating process described by Chang [10]. McClimans argues that a necessary requirement for the ongoing coordination of a construct with a method of measurement is that both are informed by the right ethical and social values, in addition to the usual epistemic considerations.

Which ethical and social values are 'right' depends on the purpose of the measurement, and on the interests and values of interested parties. In a recent paper, Alexandrova and Fabian [22] describe a democratic process for conceptualizing 'thick concepts' that are the usual concern of psychosocial measurands. Thick concepts are those that are also essentially evaluative. These deliberations are undertaken by experts that are representative of interested parties ranging from philosophers with theoretical expertise of the relevant concept, to client groups with expertise acquired through their lived experience. They propose the best way to characterize such concepts is through a co-creative process that involves the relevant interested parties in a legitimate political process of deliberation. On their view, this co-creative process is the best answer for how to manage value judgments; otherwise, they are made implicitly or without discussion by researchers designing a measure.

Building on these insights, this article proposes the notion of *ethical iterations*. Ethical iterations are structured procedures for implementing the value deliberations involved in psychosocial measure design and use. They are structured to ensure that such deliberations are comprehensive, explicit, participatory, and evidence-based, and to improve the fitness of a measurement practice to its purpose. Ethical iterations are comprehensive in two respects. First, their scope – the thing being iterated – is an entire measurement *practice*, including the theories, instruments, administration procedures, interpretive procedures, data management procedures, and decision-making procedures surrounding measurement in a given context. For example, measuring depression in a youth health service centre involves more than just a questionnaire. A set of assumptions and practices are involved in selecting which questionnaire to use, administering it, interpreting its scores, discussing the scores with respondents, recording and storing response data, and choosing a plan of action based on the data. It is this entire practice that is the object of value deliberation under ethical iterations. Second, ethical iterations are also diachronically comprehensive, encompassing all stages in the life of a measure from design to use, including any subsequent modification, translation, and repurposing of the measure. Ethical iterations can therefore be viewed as a general schema for implementing McClimans' 'epistemic dialogue' and Alexandrova and Fabian's 'legitimate political process'. At the same time, ethical iterations also produce a normative justification for the claim that a measurement practice generates knowledge about the intended construct in a given context. Ethical iterations are therefore part of the validation procedure as understood by Messick, Kane and Shepard, and complement the need for epistemic iterations. Before presenting the steps of ethical iteration in detail, we turn to discussing a case study that illustrates the centrality of ethical and social values to psychosocial measurement practice, and motivates the need for ethical iterations.

5. Case study: Measuring youth mental health in clinical settings

Our study is based on a collaboration between philosophers, health outcome researchers and service providers at Foundry, a network of integrated health and social service centres for young people aged 12–24 years across British Columbia, Canada. The first stage of the project ran from September 2019 to August 2021. It was a knowledge exchange project between researchers and service providers. In it, we developed and delivered a new training for Foundry service providers on the conceptual and practical aspects of youth mental health measurement. The training was based on insights from the philosophy of measurement, from psychometric methodology, and from the hands-on experience of our collaborators who work with youth.

The project proceeded in two phases: in Phase I, from September 2019 to February 2020, we reviewed the psychometric literature on mental health measurement and mapped measurement practices at Foundry. We compiled a draft guidebook on measurement for youth mental health clinicians, and held a workshop at McGill University to which we invited youth mental health advocates, Foundry leadership, health outcome measurement experts, and ethicists, who provided feedback on the draft. During Phase II, from February 2020 to August 2021, we heavily revised the guidebook based on feedback from workshop participants, produced a two-day training course and 11 training videos for Foundry clinicians based on the revised guidebook, and piloted the training remotely with 21 Foundry clinicians. We are currently preparing the guidebook for publication and refining our training program based on feedback from youth and healthcare

[12] Messick and Shepard's notions of validity are significantly broader than the notion of construct validity developed by Cronbach and Meehl [32]. Cronbach and Meehl limit construct validation to the process of identifying a construct that can reasonably account for observed variation in test scores. A consequence of our argument in this article is that strictly epistemic criteria of this sort are insufficient to secure the validity of a psychosocial measuring instrument.

professionals. For example, in September 2022 we held a daylong workshop on measurement at the conference of the International Association for Youth Mental Health (IAYMH).

The process of designing the course, writing its training materials, and delivering it to service providers was a fruitful opportunity to learn and think about the way Foundry uses mental health questionnaires, and the challenges Foundry faces in integrating measurement into clinical practice. At the time of writing, Foundry is a network of 16 centres, with a further 19 in development and a virtual care team that together serve about ten thousand youth per year. It follows an integrated service model, sometimes called a 'one-stop shop', where young people can access a variety of services and support in a single place, including mental health, substance use, physical and sexual health, social services, and peer support. We will focus our discussion on the use of mental health questionnaires such as the nine item Patient Health Questionnaire (PHQ-9) [41] at youth mental health clinics. Its items are based on the nine symptom categories in the DSM-IV[13] for the diagnosis of a Major Depressive Episode. The PHQ-9 asks respondents to rate the frequency at which they experienced various problems over the past two weeks, such as having poor appetite, having trouble concentrating, and feeling hopeless. Each response option is associated with a number, and the nine responses are summed up to obtain a total score.

At Foundry, health questionnaires are used for primarily four purposes. Namely, for screening clients,[14] assisting in diagnosis, tracking client progress, and evaluating Foundry's overall impact on the health of its target population. Our conversations among the research team and with service providers revealed three gaps that make the integration of measurement into clinical practice especially challenging:

1. There is a disconnect between psychometric literature and clinical practice;
2. There is a tension between the aims of data collection and patient-centred care; and
3. There is no systematic guide on the ethics of mental health measurement in a clinical context.

First the disconnect. Psychometrician Stefan Cano has lamented the lack of methods with which one can evaluate, "…the quality of the many thousands of measurement instruments in the social sciences" [42], p. 2. This resonates closely with our conversations with service providers. Often, service providers have multiple measurement instruments for what is supposedly the same construct, with little guidance for how to choose among them. Different measures exhibit varying sensitivity across the severity range of a given construct. Many measures are not developed for a service delivery context, but for use in national surveys or clinical trials. The Kessler Psychological Distress Scale (K-10), for example, has been used by Foundry and other integrated youth services to screen new clients and assign them to initial consultations, although this tool was designed to assess the nationwide prevalence of mental illness as part of the US National Health Interview Survey [43].

Moreover, most of the established mental health measures on offer were not developed specifically for youth or young people. The original PHQ-9, for example, asks respondents if they have difficulty concentrating while reading a newspaper, an activity that contemporary teenagers are unlikely to be familiar with. More importantly, insufficient tailoring of measures for this age group risks ignoring aspects of mental health that are important to young people, and over-emphasizing aspects that they consider less important.

This last point ties into the second gap; the tension between the aims of measurement and patient-centred care. The service providers we spoke with frequently expressed doubt that the information gathered through health questionnaires was beneficial for therapeutic purposes. Much of the time, they reported feeling like standardized measures were merely data collection tools rather than a clinical tool that could provide helpful information to enrich their interactions with clients. They pointed out that the time used administering and interpreting scales like patient reported outcome measures (PROMs) might be better used for genuine conversations with clients. Moreover, the questionnaires are viewed as problem-focused. This makes them clash with the strength-based and solution-focused therapy that Foundry service providers tend to offer. On the other hand, Foundry implements an evidence-based approach to service delivery that mandates the use of self-reported measures at various points along a client's health journey. Measurement is promoted at Foundry as a means of ensuring efficient and equitable care and of making youth voices heard. Foundry administrators also view measurement as important for demonstrating the overall impact of the organization, e.g., for reporting to funders. Hence, while there seem to be clear incentives for the organization at large to collect data on those accessing their services, and to have some notion of its impact at a population level, there remains a discomfort with the use of measurement in clinical situations.

The perceived lack of fitness of established mental health questionnaires with Foundry's purposes stems in part from a misalignment between the values of interested parties (e.g., youth, parents, service providers, decision-makers, and Foundry administrators) and the values that inform measure design and use. The PHQ-9, for example, was designed and validated without explicit mention of any ethical or social values. Implicitly, its use seems to be informed by a combination of such values, including overall patient wellbeing, service provider authority (in its close adherence to the DSM), and institutional efficiency.[15] Making the PHQ-9 fit for the purposes of measurement at Foundry, we will argue, requires making such values explicit and comparing them with those of interested parties linked to Foundry. Evidence concerning the alignment between these value sets should then be used to either adjust the PHQ-9; replace it with another tool that is a better fit for Foundry purposes; or adjust the service delivery context to accommodate a dialogue that facilitates a critical engagement with the instrument. Such a dialogue will contextualize its use and facilitate a joint interpretation between the young person and the service provider for a given purpose. This would be an instance of an ethical iteration, a process we will define and clarify below.

A stumbling block on the way to the sort of normative reflection we are proposing is that, until recently, there has been little to no methodological guidance on how to pursue it. This is the third gap. While there is a sizable body of literature on the ethics of research and on data governance and privacy, there is almost no guidance for service providers on the ethics of measuring in clinical contexts.[16] The module of our training that explored these issues met with considerable enthusiasm among service providers, and suggested the need for a normative framework that service providers and health outcome researchers could use for assessing whether a measure was fit for their purposes. The next section lays out some of the conceptual groundwork for such a framework.

Key values that came up often in our discussions at Foundry were ethical values like well-being and autonomy of clients; social values like inclusion, empowerment, or collaboration; pragmatic like the efficiency and scalability of the service; and epistemic, like the accuracy and comparability of measurement results. What we found was that without evidence that the use of measures is informed by the right sort of values,

[13] This was the latest version of the DSM when the measure was initially designed.

[14] Going forward we will stick with the term 'client' as this is the preferred term for young people accessing the services provided by Foundry. The term client is a value-based choice meant to emphasize patient autonomy.

[15] The PHQ-9 was initially validated against data on respondents' clinic visits and disability days (Kroenke et al. [41]).

[16] The guidebook for clinicians mentioned above will emphasize the ethical aspects of measuring youth mental health.

and that it fosters the desired individual and social consequences, serious worries for service providers and young people remained. Most important, measurement may intervene in the lives of young people in a manner that is not helpful and targeted to their real concerns. Without attention to the right values, it is doubtful whether measurement is clinically useful even if measures have been validated against established standards. For example, measure use will be undermined if service providers continue to view measurement as a mere burden; or if the content of the instruments runs counter to the therapeutic goals of Foundry; or because the instruments fail to measure what matters to young people and service providers in their context. To address these concerns, among others, measurement must encompass normative concerns that go beyond the strictly epistemic.

To ensure a measurement practice works as intended, Foundry must balance satisfying its core values as an organization; the values of mental health researchers and psychometricians; service providers; and, most important, the values of the young people, families, and communities they serve. Lack of attention to this balancing act may result in an overly rigid measurement practice that is not free to adapt to Foundry's purposes and the evolving needs of the communities. This is indeed what we found: currently, strict adherence to using psychometric tools that are deemed 'established' from a research perspective, such as the PHQ-9 and K-10, may be limiting Foundry's ability to use measurement effectively for their purposes. The next section will link this lesson to the general philosophical considerations on values in measurement discussed earlier.

6. Coordination and values in measurement practice

Recall the four reasons stated earlier for why non-epistemic values play a legitimate role in the coordination of psychosocial measurands with psychometric instruments. First, because the purpose of measurement in the case of psychosocial measurands may be undertaken for straightforward normative ends. Second, because psychosocial measurands are *Ballung* concepts that are operationalized through choices that are guided by value considerations. Third, because measurement in the psychosocial domain affects people positively or negatively; there are ethical consequences to measurement. And fourth, because the theory that underpins our understanding of response behaviour to questionnaires is underdeveloped. Each of these are exemplified in the case of Foundry.

In the first instance, measurement is implemented at Foundry both to assist the organization in facilitating the delivery of services to each young person, and in order to understand how resources are being used and where they may be best deployed upon review. In other words, measurement at Foundry has the straightforward normative end of intervening positively in the services that are delivered to each young person and to communities at large, with the ultimate aim of improving their health and well-being.

This leads us to consider the second reason. In the case of psychometric tools for mental health, just *what* is being measured is constituted by the social and institutional context that underpins how interested parties understand mental distress. Many psychometric scales for mental health are developed with DSM disorder categories as their original reference. However, as Nancy Andreasen [44], p. 111, points out, the DSM-III[17] itself is explicit in stating that:

> DSM-III provides specific diagnostic criteria as guides for making each diagnosis since such criteria enhance interjudge reliability. It should be understood, however, that for most of the categories the diagnostic criteria are based on clinical judgment, and have not yet been fully validated by data about such important correlates as clinical course, outcome, family history, and treatment response. Undoubtedly, with further study the criteria for many of the categories will be revised [47], p. 8.

There is little agreement over the right etiology of mental illness categories, and the validation of DSM categories is an ongoing project. Embedded in the development and use of tools grounded on the DSM is a respect for its authority along with the normative dimensions of dysfunction in the DSM itself. As Dominic Murphy [45] explains, claims about a 'disorder' in psychiatry usually follow a two-step process. In the first instance, there is a descriptive claim about some phenomenon, and in the second step there is a normative claim about the desirability of the phenomenon itself for the patient. Murphy discusses 'constructivist' and 'objectivist' interpretations of this two-stage picture of psychiatry. At issue is the scope of what counts as normative and what counts as merely descriptive. On the objectivist interpretation, there can be a dysfunction, or illness, and the normative aspect of the claim is whether one thinks this dysfunction is harmful. In a constructivist picture, the idea of dysfunction is itself value-laden. One need not decide between these accounts of psychiatric distress here. Regardless of which side one falls on in this debate, there is an inescapable role for values in deciding what phenomena researchers and service providers attend to as pathological.

While the nature and boundaries of the constructs at issue are in dispute, health outcome researchers have proceeded to build scales around them anyway. While this need not be problematic, it signals that the operationalization of distress as depression may or may not cohere with the way respondents experience or understand their distress. This may be due to a variety of factors, including because the questions may be crafted for a different age group, or because, as it is often the case at Foundry, the respondents may be from diverse communities, where a different understanding of distress may be operative. In addition, Foundry's own values guide the therapeutic goals they seek with the young people that they serve. As mentioned earlier, one key issue that was mentioned often was the desire to implement a 'strength-based' model of care, but often having to use measurement instruments that were problem focused. As the Truijens et al. [24] study makes clear, item content and wording can shape the symptom experience and focus of respondents on those areas suggested as important by the questionnaire. Items that ask about deficits in function, say, versus how much a person has been able to achieve, will have a different impact on respondents.

This leads us to consider the third reason, the effect of measurement on respondents. For instance, items may elicit sensitive information that needs to be dealt with appropriately as a matter of the measurement context. For example, the ninth item of the PHQ-9 asks about suicide. It would be important for a young person to be able to discuss their response with a service provider after they complete the questionnaire. Eliciting responses that may recall traumatic events or precipitate harmful behaviour requires a suitable service infrastructure for processing this information with the young person. Respondents may also interact positively with measures, finding language that helps them articulate and contextualize their distress in ways that are therapeutically helpful [24]. Much of how respondents receive and experience their interaction with a measure depends not only in the content of the measure itself (e.g. through the careful wording of items), but also on the context within which the person interacts with the instrument. This includes the space in which the measure is used, and the work done by a service provider to contextualize the purpose of the tool and to debrief a client on how they understood what the items of the measure queried. In this way, attention to values like empowerment or autonomy directly impacts the interpretation of the tool. Attention to such values is required for the ongoing successful coordination of the tool in the shifting context of its use and the people using the measure.

This includes the trust and confidence clients feel when asked to share information about themselves that will be recorded and stored. In

[17] The most recent version is DSM-5-TR, but the same caveats listed by Andreasen still apply.

the context of Foundry, a key consideration is data ownership, control, access, and possession (OCAP®)[18] for the diverse communities across British Columbia, including Indigenous communities. Throughout Canada's colonial history and context there have been ongoing problems with the imposition of external values and classification systems on First Nations, Inuit, and Métis peoples. Indigenous communities have long alerted and highlighted concern about the collection, storage, use, and ownership of information about themselves so that it may not be misused or appropriated by others. For example, demographic and personal data gathered from communities and stored by the Canadian government facilitated the forced removal of children from their families to residential schools and the infamous '60 s scoop' [46], p. 50. Residential schools were Church and government run facilities developed to implement cultural assimilation for Indigenous children and youth, where thousands of children died. In the 60′s scoop, children were forcibly removed from their families by the government and placed for adoption elsewhere across Canada and even abroad. Overall, the misuse of data and lack of consideration for how it is owned, controlled, accessed and possessed has harmed many Indigenous communities in tragic ways. Attention to these issues is required to ensure an ethical measurement practice that serves its intended purpose. In the case of Indigenous groups, the autonomy not just of the client, but of entire communities may very much weigh heavily on a person who is asked to provide information about themselves through an instrument such as a psychometric measure.

A worry linked to the above that arises in the interpretation of measures in relation to their context and purpose of use is stigma and the possible reification of measurement categories. Measurement is a kind of labeling that can affect how people feel about themselves, and how others treat them [16,24]. Reification is a potential effect of measurement, which can be more or less harmful depending on the context of use. In the context of mental health measurement reification operates on at least two levels. In the first instance, the use of a category from the DSM endorses it as the right or preferred operationalization of an experience of distress. Risk of this 'categorical' kind of reification exists even when a client's information is presented to them in non-numerical form, e.g., using nominal or ordinal scales. In the second instance, the use of a measurement instrument presents the attribute as quantitative and represented on an interval or ratio scale. If a young person arrives feeling distressed and takes a screening test that tells them they have 18 (or some other number) of depression, this may reinforce an illness category in a specific way. As we have stated, measurement can be an authoritative scientific practice, and scores, most often expressed as numbers, can be perceived as authoritative entities. If someone perceives a test score as numerical confirmation that they are ill, this may harm how they see themselves and how others see them. This is more so the case if at their first interaction with the centre, their initial impression is that they have some amount of some real attribute. This may be the desired interpretation, but it may not. In the case of screening, the inference supported by the score provides evidence for a specific decision about where a young person may best be served in the care centre.

Finally, the fourth reason is a lack of clear theory that guides the interpretation of differences in response behaviour, including differences in the same young person over time. In the absence of detailed insight into the dynamics of response behaviour, attention to the context and purpose of use is essential in order to interpret any given score. The appearance of psychometric scales as interval scales that may function much like a ruler, or that can be interpreted much like a thermometer, may be misleading. Proceeding as if one has an accurate theory of response behaviour runs the risk of overemphasizing changes in scores, or of suggesting that scores map neatly into therapeutic outcomes. In particular, in the absence of such a theory, service providers are less warranted in interpreting different scores over time as legitimate changes in degrees in the construct, or to attribute them to the quality of a therapeutic intervention. This difficulty in standardization, for instance, is what leads McClimans [40] to endorse a dialogue model that will enable genuine conversations between service providers and respondents. Such conversations allow both sides to better understand how to interpret the significance of scores for each individual respondent. This leaves room for a wide set of non-standardized judgments to take place in the interpretation of scores for the purposes they are intended, within the given context of use. These judgments are informed in each case by the organizational, community and personal values of each party involved in the measurement practice.

One could at this point object that ethical considerations are important for improving other virtues of a measure, such as safety, social acceptability, ease of use, or relevance for the target population, among others, but have nothing to do with its authority, which remains purely epistemic. This echoes the point alluded to in the first section; normative considerations about the use of measurement instruments are inherent in any context of use. The comfort of a child while their fever is measured is an important normative dimension for the use of a measurement instrument, but it is not related to the authority of the measurement outcome. The distinction we are pressing is between temperature as a measurand versus something like depression, where normative considerations are pervasive. If the temperature measured is that of a bucket of water, then ethical concerns in relation to the kind of measuring instrument are dissolved, but the authority of the measurement is unaffected. However, whether a client feels reflected in the items of an instrument, and whether the instrument has a meaningful therapeutic interpretation that is useful for that context and young person will depend on the information the items ask for *and* on the legitimacy of the values that inform the context of use, the interaction, and the items themselves.

We argue that psychosocial measurement is distinct because alignment with the ethical and social values of interested parties is constitutive of the *very ability of measurement to produce authoritative knowledge claims*. For the reasons already described, ethical and social value judgments are necessary for justifying the claim that a psychosocial measurement practice is fit for its purpose, namely, that it measures the right construct for a given context and population, and that it measures it well. We therefore resist the assumption that it is possible to validate a psychosocial measuring instrument independently of the ethical and social values that guide its use.

For example, suppose that the developers of a hypothetical tool that uses brain imaging to measure depression tried to assess its validity.[19] It is tempting to try to assess the validity of the tool in strictly epistemic terms, analogously to how one would establish that an instrument measures temperature. Much like a thermometer, the developers of the brain imaging tool would have to coordinate a concept of the measurand – in this case, depression – with their tool. This coordination would establish a mapping between patterns of data representing brain activity and levels of depression. Yet 'depression' has multiple meanings, and under each meaning, individuals or groups of people would be evaluated differently, leading to different individual or social goods and harms. In this respect, depression is unlike temperature, whose various meanings (e.g., kinetic vs. thermodynamic) are not associated with distinct individual or social goods or harms.[20]

In choosing which concept of depression to coordinate with their new tool, the developers would have to decide, either explicitly or implicitly, which norms govern the evaluation of individuals or groups of people with respect to depression, and then judge the validity of their

[18] OCAP® is a registered trademark of the First Nations Information Governance Centre (FNIGC). https://fnigc.ca/ocap-training/.

[19] We thank one of the anonymous reviewers of this paper for this example.

[20] We are not claiming that this is the case for all physical measurands. For example, measuring global warming is likely to involve similar ethical and social value judgments as the measurement of psychosocial measurands.

tool against these norms. Such norms may be, for example, etiological classification, prognostic classification, counting of symptoms, severity of symptoms, impact on functioning, or some combination of these. In choosing an evaluative norm, the developers would have to make an ethical or social value judgment about what counts as a good measurement of depression. The 'good' here is no longer merely epistemic, but involves a choice as to which individual or social goods or harms the evaluation is intended to promote or mitigate. This value judgment then becomes constitutive of the judgments developers make about the validity of the new tool.

The justification for choosing one norm of evaluation over another must be provided partially in terms of some individual or social goods or harms that can be reasonably expected to result from evaluating individuals or groups under that norm. Whether or not the new tool is indeed valid therefore depends, at least partially, on whether its use promotes the attainment of those goods and the mitigation of those harms. This can only be tested by using the tool in real-life settings. The success of the tool in promoting the specified goods and mitigating the specified harms depends on more than just the properties of the tool in isolation. Rather, it also depends on whether the tool is used and interpreted in ways that align with the goals and values of other interested parties, such as patients, clinicians, researchers, and policymakers. Validation is therefore both value-laden and context-specific.

For the same reasons, when a psychosocial measurement practice is misaligned with the aims and values of interested parties, they are justified in doubting or rejecting the authority of measurement outcomes, and not just the relevance or usefulness of those outcomes. Foundry is a case in point: clinicians are justified in doubting that many standard mental health questionnaires, which were designed and validated in light of different values than their own, measure mental health as understood by their client population and as mandated by Foundry's mission. Of course, the same questionnaires, when embedded in a different measurement practice that is geared toward a different purpose and population, may measure a mental health construct well. If they do, it would be partially thanks to an alignment between the non-epistemic values informing the design and use of the questionnaires and those of interested parties in the alternative context. Yet to think that there is a value-free fact of the matter as to whether and how well a questionnaire measures a psychosocial construct is precisely the mistake we have argued against. The next section describes a process through which value alignment may be improved, and with it the coordination between the construct and measurement practice in a given context.

7. Ethical iterations

The reasons in the previous sections suggest that a framework to make these social and ethical criteria explicit in the measurement practice, and to link them to epistemic criteria, is needed in order to secure the authority of psychosocial measurement. We follow the iterative framework developed by van Fraassen [11] and Chang [10] and suggest concurrent ethical iterations are necessary to coordinate a concept in the behavioural and social sciences with its methods of measurement alongside epistemic iterations. Ethical iterations have the same aim as that articulated by Chang [10]: to ensure progress in the coordination of measurement instruments and their measurands. As he explains,

> There are two modes of progress enabled by iteration: *enrichment*, in which the initially affirmed system is not negated but refined, resulting in the enhancement of some of its epistemic virtues; and *self-correction*, in which the initially affirmed system is actually altered in its content as a result of inquiry based on itself. Enrichment and self-correction often occur simultaneously in one iterative process… [10], p. 228.

Without attention to ethical and social values, psychosocial measurands will remain under-specified, and their coordination will be incomplete. This is because the enrichment of the measurands and their attendant measurement practice depends on understanding the normative ends for which we undertake any measurement; the values and social practices which shape the measurand; the impact of measurement in a given context; and the value judgments necessary to successfully interpret measures in those contexts. Similarly, the possibility of self-correction is foreclosed if these value considerations are not made explicit and cannot be deliberated upon by all interested parties in an inclusive process. For these reasons, ethical iterations must be undertaken alongside epistemic iterations in order to allow the measured construct and measurement procedure to be informed by sound ethical and social values that are a good fit for the intended purpose of the measure and beneficial to interested parties.[21]

By 'ethical iterations' we mean the intentional repeated process of (i) reflecting on the values that guide current theorizing about a construct and the current design, selection, and/or interpretation of a measure of that construct; (ii) collecting evidence about the consequences of theorizing about the construct, and of designing, using, and/or interpreting the measure, in light of these values; (iii) collecting evidence about the values and concerns of interested parties affected by using the measure for a given purpose; (iv) revising values, their relative importance, and their trade-offs in light of evidence as well as general normative considerations; and (v) modifying conceptions of the construct and the design, selection, and/or interpretation of the measure in light of the previous steps. Fig. 1 summarizes these five steps.

At Foundry, the ongoing ethical scrutiny of a tool like the PHQ-9 would require an iterative process of, (i) reflecting on the values that guide the design of the PHQ-9 and the conception of depression it presupposes, i.e., the DSM concept of major depressive disorder; (ii) collecting evidence about the effects of using the PHQ-9 for screening in light of these values; (iii) engaging young people and their communities, as well as service providers and Foundry management, about what matters to them as clients and service providers; (iv) comparing, revising, or elaborating a set of values based on Foundry's values and those of young people and their communities, as well as on general ethical principles or frameworks; (v) deciding whether a construct like depression is the right target, and whether the PHQ-9 is the right tool, depending on the purpose of the measure and on Foundry's values. This process would be repeated regularly, either at fixed time points or as the need arises.

For example, in this last step, interested parties can decide that a pathology-based construct like DSM depression is not the best target for initial screening at their centres; they could decide to eschew the PHQ-9 for a different measure; or to select only specific items from the instrument that might be most useful and are not contrary to their screening aims and other values, like youth empowerment; they may decide to change the wording or response options of some of the items; to share background information with clients prior to handing them the screening instrument; or to encourage a conversation between client and service providers about the responses the young person gave to items in the PHQ-9 or other measure. This list does not exhaust the available options, and it illustrates that the changes that happen in light of ethical iteration do not just concern the content or wording of the instrument itself. We have made sure to describe measurement as a practice for this reason. The purposes, context and interested parties involved in a

[21] General normative considerations, such as appeal to theoretical frameworks on well-being or medical ethics, are a necessary component of the deliberation we envision here. In this paper we make the simplifying assumption that the values of interested parties can ultimately be reconciled with each other and with general normative considerations. Further work, and in particular the normative theory of measurement we mention in the final section, is required to address cases of persistent value conflicts and the possibility of mistaken values from interested parties.

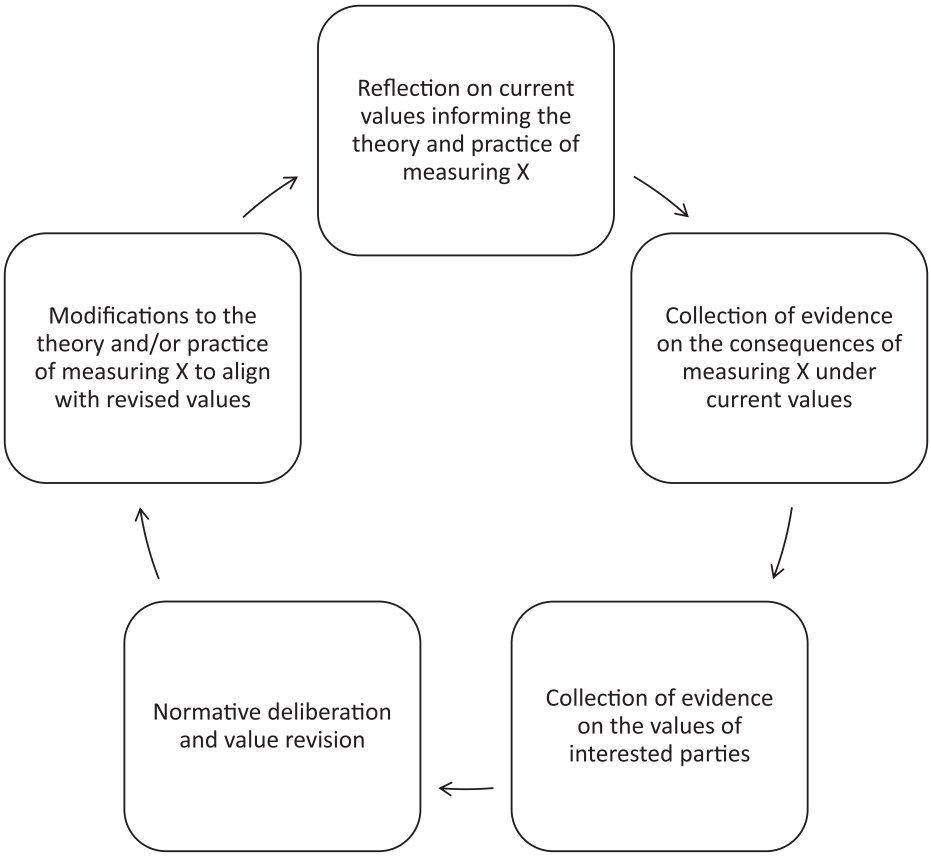

Fig. 1. The framework of ethical iterations consists of five steps that progressively align measurement practice with the values of interested parties in an evidence-based manner.

measurement practice are constantly changing, and ethical iterations reconfigure measurement practice as a whole to ensure that it is ethically justified.

The above illustrates the case for a tool that has already been designed, where ethical iterations would guide the use, interpretation, and inferences that are supported by the values that are identified in this kind of reflective exercise. It is worth stressing the need to think carefully of the interlocutors in this exercise. It should broadly include the interested parties that will undertake or be affected by the use of the tool and be part of the measurement practice. Ideally, however, ethical iterations should be undertaken from the very beginning of a design process. In other work currently under preparation, we trace the ethical iterations intrinsic to measure design, and argue that they should be made explicit in order to justify the fitness of the measurement practices they serve to their intended purposes. It is important to stress that we are not proposing that values be introduced into psychosocial measurement. We hope to have shown that non-epistemic values are already intrinsic to these measurands and instruments, and what is required is a framework to make these choices rigorous and intelligible for interested parties.

Our example reveals the importance of ethical iterations and why they must also occur at the use stage of a measurement practice. At Foundry, it was identified early that service providers and clients must be part of the deliberative process that informs which tools to use, how to use them, and for what purposes. Yet there was no established procedure at Foundry for carrying out such deliberations. The choices in these deliberations should happen in light of the values put forward by the different interested parties – e.g. the organization, youth, and their communities. Otherwise, the interpretation and use of measures in that context will not measure what matters to interested parties, and will not measure in a way that promotes other interested parties' values, like empowerment. This may arise if service providers are expected to apply existing measures without adjusting their measurement practice to their values, those of their organization, and other key interested parties like young people and members of their communities. It may also arise because the processes for addressing what matters to interested parties from a values perspective is at the initial stages of implementation or lacks an explicit procedure. To secure their measurement goals, there was increasing recognition by Foundry researchers and service providers for the need to align the values of the organization and the communities they serve with a process for contextualizing and interpreting measures. This was important given the organization and other interested parties were not involved in developing the measures they were using. Partly in light of this, a procedure is required that ensures the measures selected and the process of their use aligns with their own ethical and non-epistemic value commitments. Even if they had created their own measures, however, a procedure to ensure the ongoing legitimacy of the measure is required to ensure its ongoing fit.

The normative, reflective exercise of ethical iterations loops back and informs the theorizing and operationalization of the construct in the next iteration. Again, ethical iterations are already implicitly undertaken when psychometric measures are designed and used, but they are usually not recognized as such, and are often undertaken in a partial and suboptimal way that does not sufficiently acknowledge different parties' interests, differences across populations, contexts, and uses, and the complex balancing among competing values. The case of Foundry illustrates the importance of making ethical iterations explicit for selecting and using measures for youth mental health in a clinical context. Such iterations, combined with epistemic iterations provide joint grounds for the productive ongoing coordination of psychosocial measurands and their instruments. In particular, ethical criteria of justification ensure that the enrichment and self-correction Chang [10]

has in mind as necessary for a progressive coordination practice happens in the case of psychosocial measurands as it should. This iterative process begins at the design stage of a measure, ideally by integrating interested parties and experts in a productive deliberative process [22]. As our work shows, the values and meaning embedded in the instrument need to be negotiated in an ongoing way at the use stage of the instrument as well. The notion of ethical iterations captures the structured deliberative processes in the long arc of the life of the instrument that may allow it to be coordinated for different purposes and contexts of use.

8. Toward a normative theory of measurement

We have argued that non-epistemic values play a legitimate role in coordinating psychosocial measurands with a measurement practice, thereby securing the authority of measurement. We provided four reasons why non-epistemic values can play this role and illustrated these reasons for the case of youth mental health measurement at Foundry. Despite the centrality of non-epistemic values to psychosocial measurement, their importance is rarely fully acknowledged. Even when they are viewed as relevant to measurement, they are often mistakenly relegated to mere safeguards against harmful consequences. By contrast, we have argued that ethical and social values play a necessary role in justifying the claim that a psychometric instrument measures what matters to interested parties. We then proposed a general structure for a procedure – called 'ethical iteration'– that helps make explicit the value considerations at the heart of psychosocial measurement practice. A measurement practice that incorporates regular ethical iterations is progressive, in the sense that it is sensitive to the varied and changing aims and values of respondents, decision makers, researchers, and other interested parties. Although we focused on the example of mental health, our proposal is meant to apply broadly to measurement across the behavioural and social sciences. Variations in the implementation of ethical iterations across these fields, and field-specific challenges involved in their implementation, are topics for future research.

The foregoing discussion suggests that a normative theory of measurement is required that would link epistemic justifications with ethical ones. The first question it should answer is whether measuring is the right thing to do in a given case. We pose this question as a normative one. Measurement is an act that has consequences for the lives of humans, and as such requires ethical justification. The second question is what and how researchers and service providers ought to measure in a given set of circumstances. Again, we intend the 'ought' in an ethical sense. Third, it would be important to specify under what conditions measurement-related actions, such as score interpretation, the application of specific methods of data analysis, and measurement-based decision-making, are justified. Epistemic justifications are surely relevant for answering such questions: a measure should not be used to make decisions about medical treatment if it is not sensitive enough. But epistemic justifications do not exhaust the answer, for the reasons already mentioned above.

Such theory cannot be purely formal, like some theories of inductive risk, but must be sensitive to the interests and values of different parties. Specifically, a normative theory of measurement should provide standards of evidence for when measurement is permissible and beneficial to various interested parties. Such standards would then guide the evidence-collection steps (ii) and (iii) of ethical iterations, as detailed above. The current absence of such standards, and the ensuing lack of clarity as to what counts as good evidence for the fitness of a measure to its purpose, explain the three gaps we identified above.[22] The joint iteration of epistemic and ethical criteria has the potential to improve the fitness for purpose of a psychosocial measurement practice. We take fitness-for-purpose to be a thoroughly normative notion. Determining the fitness-for-purpose of a measure involves making ethical judgments concerning why, what, and how behavioural and social scientists *ought* to measure in a given context. We plan to develop the notion of fitness for purpose for psychometrics in future work.

[22] They are also potential causes behind the limited practical adoption of Messick and Kane's argument-based approach to validity.

CRediT authorship contribution statement

Sebastian Rodriguez Duque: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Eran Tal:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Skye Pamela Barbic:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

We would like to thank Danielle Celone, Kyle Dewsnap, Sophie Osiecki, and Darius Valevicius for their help in developing the measurement training that inspired some of this work. This work was supported with funding from the Canada Research Chairs program (grant CRC-2019-001199) and the Canada First Research Excellence Fund (CFREF) via the Healthy Brains for Healthy Lives knowledge mobilization grant and funding from CIHR Grant #w12-179949.

References

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders Fifth Edition Text Revision DSM-5-TR. 2022.
[2] A. E. R. Association, *Standards for educational and psychological testing*. American Educational Research Association, 2014.
[3] M.T. Kane, An argument-based approach to validity, Psychol. Bull. 112 (3) (1992) 527–535, https://doi.org/10.1037/0033-2909.112.3.527.
[4] S. Messick, Test validity and the ethics of assessment, Am. Psychol. 35 (11) (1980) 1012–1027, https://doi.org/10.1037/0003-066X.35.11.1012.
[5] L.A. Shepard, Chapter 9: Evaluating test validity, Rev. Res. Educ. 19 (1) (1993) 405–450.
[6] L.A. Shepard, Evaluating test validity: Reprise and progress, Assess. Educ. Princ. Policy Pract. 23 (2) (2016) 268–280.
[7] K.C. Elliott, Values in Science, Elem. Philos. Sci. (2022), https://doi.org/10.1017/9781009052597.
[8] P. Rooney, On values in science: Is the epistemic/non-epistemic distinction useful?, in: PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association Cambridge University Press, 1992, pp. 13–22.
[9] L. Mari, M. Wilson, A. Maul, Measurement across the Sciences, Springer, 2021.
[10] H. Chang, Inventing Temperature: Measurement and Scientific Progress. Oxford University Press, 2004. Accessed: May 24, 2021. [Online]. Available: http://oxford.universitypressscholarship.com/view/10.1093/0195171276.001.0001/acprof-9780195171273.
[11] B.C. van Fraassen, in: B.C. van Fraassen (Ed.), Scientific Representation: Paradoxes of Perspective, Oxford University Press, 2008, https://doi.org/10.1093/acprof:oso/9780199278220.003.0006.
[12] M.T. Kane, Validating the Interpretations and Uses of Test Scores, J. Educ. Meas. 50 (1) (2013) 1–73, https://doi.org/10.1111/jedm.12000.
[13] C. Larroulet Philippi, Valid for what? On the very idea of unconditional validity, Philos. Soc. Sci. 51 (2) (2021) 151–175.
[14] I. Hacking, The Taming of Chance, Cambridge University Press, 1990.
[15] T.M. Porter, Trust in numbers: the pursuit of objectivity in science and public life, Princeton University Press (1996), https://doi.org/10.1515/9781400821617.
[16] S. Gould, The Mismeasure of Man, WW Norton & Company (1996).
[17] L.D. Wijsen, D. Borsboom, A. Alexandrova, Values in psychometrics, Perspect. Psychol. Sci. 17 (3) (2022) 788–804.
[18] N. Cartwright, R. Runhardt, Measurement, in: N. Cartwright and E. Montuschi, (eds.), *Philosophy of social science: A new introduction*, OUP UK, 2014, pp. 265–287.

[19] D.M. Hausman, Valuing health: Well-being, freedom, and suffering, Oxford University Press, 2015.
[20] L. McClimans, J. Browne, S. Cano, Clinical outcome measurement: Models, theory, psychometrics and practice, Stud. Hist. Philos. Sci. Part A 65–66 (2017) 67–73, https://doi.org/10.1016/j.shpsa.2017.06.004.
[21] S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, Am. Psychol. 50 (9) (1995) 741, https://doi.org/10.1037/0003-066X.50.9.741.
[22] A. Alexandrova, M. Fabian, Democratising measurement: or why thick concepts call for coproduction, Eur. J. Philos. Sci. 12 (1) (2021) 1–23, https://doi.org/10.1007/s13194-021-00437-7.
[23] I. Hacking, The Social Construction of What? Harvard University Press, 1999.
[24] F.L. Truijens, K. Van Nieuwenhove, M.M. De Smet, M. Desmet, R. Meganck, How questionnaires shape experienced symptoms. A qualitative case comparison study of questionnaire administration in psychotherapy research, Qual. Res. Psychol. 19 (3) (2022) 806–830, https://doi.org/10.1080/14780887.2021.1886383.
[25] Joint Committee for Guides in Metrology, JCGM, Evaluation of measurement data - Guide to the expression of uncertainty in measurement, 2008.
[26] E. Tal, How Accurate Is the Standard Second? Philos. Sci. 78 (5) (2011) 1082–1096, https://doi.org/10.1086/662268.
[27] D. Borsboom, Measuring the Mind: Conceptual Issues in Contemporary Psychometrics, Cambridge University Press, 2005.
[28] M.A.G. Sprangers, C.E. Schwartz, Integrating response shift into health-related quality of life research: a theoretical model, Soc. Sci. Med. 48 (11) (1999) 1507–1515, https://doi.org/10.1016/S0277-9536(99)00045-3.
[29] A. Vanier, B. Falissard, V. Sébille, and J.-B. Hardouin, The complexity of interpreting changes observed over time in health-related quality of life: A short overview of 15 years of research on response shift theory, in: *Perceived Health and Adaptation in Chronic Disease*, Routledge, 2017.
[30] L. McClimans, Patient-Centred Measurement: Ethics, Epistemology, and Dialogue in Contemporary Medicine. In progress.
[31] S. Messick, Test validity: A matter of consequence, Soc. Indic. Res. 45 (1) (1998) 35–44, https://doi.org/10.1023/A:1006964925094.
[32] L.J. Cronbach, P.E. Meehl, Construct validity in psychological tests, Psychol. Bull. 52 (4) (1955) 281.
[33] H. E. Longino, Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy, in: L.H. Nelson and J. Nelson, (eds.), *Feminism, Science, and the Philosophy of Science*, Synthese Library. Dordrecht: Springer Netherlands, 1996, pp. 39–58. doi: 10.1007/978-94-009-1742-2_3.
[34] H. Douglas, Inductive Risk and Values in Science, Philos. Sci. 67 (4) (2000) 559–579.
[35] J.L. Cerezo, Human nature as social order: A hundred years of psychometrics, J. Soc. Biol. Struct. 14 (4) (1991) 409–434.
[36] B. Evans, B. Waites, IQ and mental testing: An unnatural science and its social history, Macmillan International Higher Education, 1981.
[37] L.V. Jones, D. Thissen, 1 A History and Overview of Psychometrics, Handb. Stat. 26 (2006) 1–27.
[38] M.M. Sokal, Psychological testing and American society 1890-1930, in: This volume had its origins in a symposium of the same title, organized and chaired by Michael M. Sokal, at the 150th National Meeting of the American Association for the Advancement of Science, held in New York on May 29, *1984.*, Rutgers University Press, 1987.
[39] A. Alexandrova, A Philosophy for the Science of Well-being, Oxford University Press, New York, NY, 2017.
[40] L. McClimans, A theoretical framework for patient-reported outcome measures, Theor. Med. Bioeth. 31 (3) (2010) 225–240.
[41] K. Kroenke, R.L. Spitzer, J.B. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (9) (2001) 606–613, https://doi.org/10.1046/j.1525-1497.2001.016009606.x.
[42] S.J. Cano, T. Vosk, L.R. Pendrill, A.J. Stenner, On trial: the compatibility of measurement in the physical and social sciences, J. Phys. Conf. Ser. 772 (2016), 012025, https://doi.org/10.1088/1742-6596/772/1/012025.
[43] R. Kessler, et al., Short screening scales to monitor population prevlances and trends in non-specific psychological distress, Psychol. Med. 32 (2002) 959–976, https://doi.org/10.1017/S0033291702006074.
[44] N.C. Andreasen, DSM and the death of phenomenology in America: an example of unintended consequences, Schizophr. Bull. 33 (1) (2006) 108–112, https://doi.org/10.1093/schbul/sbl054.
[45] D. Murphy, Psychiatry in the scientific image, in: Psychiatry in the Scientific Image, MIT Press, Cambridge, MA, US, 2006, p. xi, 410.
[46] The First Nations Information Governance Centre, First Nations data sovereignty in Canada, Stat. J. IAOS 35 (1) (2019) 47–69, https://doi.org/10.3233/SJI-180478.
[47] American Psychiatric Association Committee on Nomenclature and Statistics. Diagnostic and Statistical Manual of Mental Disorders (DSM-III), American Psychiatric Association, Washington, DC, 1980.