Is Quantitative Measurement in the Human Sciences Doomed? On the Quantity Objection

Cristian Larroulet Philippi

Are widely used measurements in the human sciences (say happiness surveys or depression scales) quantitative or merely ordinal? If they are merely ordinal, could they be developed into quantitative measurements, just like in the progression from thermoscopes to thermometers? Taking inspiration from recent philosophy of measurement, some practitioners express optimism about future human science measurements. The so-called quantity objection stands out for having the only chance of settling the debate in favour of the pessimists. It claims that the problem lies not with current, or likely future, measurement practices in the human sciences, but with human science attributes themselves—they just are not quantitative, but merely ordinal. Hence, they cannot not (thus will not) be measured quantitatively. The argument has a long and distinguished pedigree. This paper assesses old and recent versions of it, namely: the objection made originally to Fechner's psychophysics by von Kries (among others) and Michell's recent version of this objection. To do so, the paper first draws important distinctions between different versions of the argument that have been overlooked. Then, it argues that none of the versions of the quantity objection provide a good reason for the optimists to give up their optimism. In particular, Michell's argument characterizes the measurand (that is, the attribute to be measured) in a way that optimists do not and need not accept. Yet the optimists' defence articulated here brings with it serious burdens to discharge. The ball is in their court.

ORCID: Larroulet Philippi, www.orcid.org/0000-0001-5793-4670

1. Introduction

Are widely used measurements in the human sciences, say happiness surveys or depression scales, quantitative or merely ordinal? Do they inform us about 'distance relations' (how much happier I am now than before) or only about ordering ones (whether I am happier)? Much research implicitly assumes that these measurements are (at least approximately) quantitative—averages are routinely computed from happiness and depression scores to infer the effectiveness of social policies and anti-depressants (Stegenga [2018]; Larroulet Philippi [2025]). Methodologists have been arguing for decades that such measurements are merely ordinal.[1] Yet how much these criticisms have or will change research practices remains unclear.

[1] For example, in medicine (Merbitz et al. [1989]), happiness economics (Bond and Lang [2019]), and psychology (Michell [1990], [1997], [1999], [2009]).

Some researchers are concerned, however, especially within psychometrics. Some authors (Stenner et al. [2013]; Bond and Fox [2015]; Bringmann and Eronen [2016]) can be read as reacting to these criticisms by developing what I will call a 'nuanced position'—while they grant that current measures are not quantitative, they remain optimistic that at least some human science attributes will end up being measured quantitatively in the future. More interestingly, recent philosophy of measurement has been deployed by these authors to buttress this nuanced position. The idea goes roughly like this: The quantification of temperature was not a straightforward task, but a hard-fought, centuries-long endeavour. As is clear from (Chang [2004]; Sherry [2011]), temperature's (approximate) quantitative measurement was not justified by some kind of one-time, straightforward verification that raw data satisfied some test for quantitativeness. Justification was gradually established, by postulating relations between temperature and other theoretical quantities, building confidence that measuring instruments were shielded from other factors (variations in atmospheric pressure, differences in the flasks used, and so on), and putting assumed-to-be-correct measurement results to the task of prediction, and (approximately) succeeding. Such endeavour took time, effort, and patience. The same things are needed for achieving success in the measurement of any other latent attribute, like reading comprehension or depression.

Some psychologists have latched onto something like this position. They insist that 'in no case [in science's history] was an attribute born quantitative' (Stenner et al. [2013], p. 2), hence, current failure is not a sign of permanent failure. They bolster their optimism by explicitly linking psychometric measurement to the history of thermometry. For example, the latest edition of a popular psychometrics textbook (Bond and Fox [2015], p. 14) encourages readers too worried about the quantitative status of psychological measurement to consult (Sherry [2011]).[2]

This nuanced position has some appeal. (After all, who can predict the future?) And something like it has been endorsed by some philosophers of science for the case of psychology (Bringmann and Eronen [2016]). Barwich and Chang ([2015]) also defend the prospects for psychophysical (sensory) measurement from methodologists' attacks precisely on the grounds

[2] Sherry did not give explicit reasons for being optimistic about quantitative human measurement. He closed his influential article thus: 'It is open to psychologists to be encouraged *or discouraged* in their attempts to quantify mental attributes by comparing their accomplishments with Black's' (Sherry [2011], p. 524).

that the challenges present in sensory measurement resemble those faced in temperature measurement.

Several methodologists remain unconvinced, however. Psychologists' reluctance to engage with specific approaches to measurement (say, conjoint measurement) and/or enough differences between the subject matters of psychology and physics make optimism unwarranted for some (Michell [1999], [2012a]; Trendler [2009], [2019]; Franz [2022]). As Michell ([2012a], p. 261) is keen to remind us: 'Just because concepts like temperature, which were first experienced only as matters of degree, were later shown to be quantitative, it does not follow that all concepts admitting degrees are quantitative'. Unsurprisingly perhaps, this debate between optimists and pessimists concerning the eventual quantification of human attributes has arguably reached an impasse.

Here enters an argument that promises to overcome the impasse, called 'the quantity objection' by psychologist and historian of psychology Edwin Boring ([1921]): The reason why human measurements are not quantitative lies not with current, or likely future, measurement practices in the human sciences. The reason lies with the attributes themselves—they are just not quantitative. Because these attributes are at most merely ordinal, they cannot—thus will not—afford a quantitative representation. The problem is metaphysical (concerning how these attributes are), not epistemic (concerning how we get access to them). So, the above optimism is misplaced.

Despite how foreign it may sound to quantitatively trained social scientists today, and how little it exercises contemporary philosophers, the quantity objection has a long and distinguished history. Section 2 briefly situates the objection historically, focusing mostly on criticisms to Gustav Fechner's psychophysical measurement, which were those that merited Boring's coining of the term. These criticisms are not historical curiosities—similar ones are seriously discussed today.

This article's contribution is twofold. First, by distinguishing different versions of the quantity objection and assessing their respective plausibility, I show that Boring's ([1921]) original rendering of the quantity objection against Fechner's measurements is problematic in two ways (sec. 3). First, it lumps the different versions of the objection into one; second, that one version happens to be untenable from today's perspectives, while others are not so, or not in the same way. The second contribution is an assessment of the more recent and compelling version of the

quantity objection: Michell's ([2009], [2011], [2012a], [2012b]). After introducing Michell's version (sec. 4), I will argue that, just like previous ones, it cannot move us past of the impasse. Michell's argument does not give a reason to optimists to give up their optimism, because the characterization of the attribute to be measured (or the 'measurand') that Michell assumes in his argument is not one that optimists accept or must accept. I will close, however, arguing that the optimists' response brings with it serious burdens to discharge. The ball is in their court.

2. The Quantity Objection: Brief Historical Background

Quantity objections are not new. In medieval physics, efforts to provide quantitative understandings of velocity or temperature were fiercely challenged:

> The conservative school supported Aristotle's principle that since quality and quantity belonged to absolutely different categories, the one could not be reduced to the other. Examples of changes in quantity were changes in length or number, which were brought about by the addition or subtraction of either continuous or discontinuous homogeneous parts. That was all the change involved. But a change in a quality such as heat was quite different. Heat might exist in different degrees of intensity, but a change in intensity was not brought about, for example, by adding one homogeneous part of heat to another. The heats of two bodies brought into contact did not make a greater heat, as the lengths of two bodies made a greater length. So Aristotle and his supporters considered that each degree of intensity of heat was a different quality, and that a change in the intensity was brought about by loss of one quality of heat and the acquisition of another. The same went for every change in quality. (Crombie [1961], p. 151)

Crombie's description of the critique resembles objections—made centuries later—targeting efforts at quantifying human attributes. In the nineteenth century, John Stuart Mill criticized Jeremy Bentham's concept of utility for assuming that pains and pleasures were homogeneous (Sen [1981]). A century earlier Thomas Reid criticized Francis Hutcheson's quantitative moral psychology similarly (Brooks and Aalto [1981]). But I will focus here on the criticisms made to Gustav Fechner's efforts at quantitatively measuring the intensity of sensations during the second half of the nineteenth century.

Edwin Boring ([1921], p. 453) coined the term 'quantity objection' to describe a particular critique made to Fechner's programme, one he summarized thus: 'Introspection […] does not show that a sensation of great magnitude ever contains other sensations of lesser magnitude in the way that a heavy weight may [supposedly] be made up of a number of smaller weights'. The

square bracket is Boring's—he did not buy into the objection. But it is worth looking beyond his summary.

Among the firsts to criticize Fechner's programme from this perspective were French mathematician Jules Tannery and the better-known Johannes von Kries. Here is Tannery (cited in Heidelberger [2004], p. 209): 'The essential characteristic of directly mensurable dimensions is *homogeneity*: whatever is added, such that something increases, is of the same exact kind as that which was already there: length, surface, and time are dimensions of this kind. If we add one length to another, both of them are of the same kind and essence and their sums are also of the same kind. Directly measurable dimensions necessarily have this quality, because measurement itself requires that dimensions of the same kind be comparable'. After this citation, Heidelberger's ([2004], p. 209) account of Tannery's criticism continues with the following: 'Tannery illustrates the lack of homogeneity in sensation with the example of the sensation of heat: If you hold an object in your hand and the heat of that object increases, at some point the threshold of pain is reached. The original sensation (heat) is of an entirely different kind than the final sensation (pain). Entirely different nerves are involved in those sensations. And what is true for these two extremely different sensations also holds for all those in between, albeit to a lesser degree'.

Von Kries thought that in the case of sensations' intensity, we just cannot conceive of the attribute as being quantitative: 'Impartial reflection leads, in my view, to the inevitable conclusion that no sense at all can be attached' to claims that presume quantitativeness, such as that this degree of intensity is much larger than this other degree, or that the difference between these two degrees is twice as large as the difference between these other two degrees (Niall [1995], p. 291). If one pictures a series of sensations of increasing intensity, he continues, then 'one must concede that there exists no quantitative relation between the different steps of a series of intensities. This is clearest in cases [like] pain. What it means to say that one pain may be exactly ten times as strong as another, is simply unfathomable' (Niall [1995], pp. 291–92). For a sense of how misguided von Kries thought the measurement of sensations was, how confident he was that his intuitions about this will be widely shared, and what he might respond to the Fechner's of today, look no further:

> Even if some people were to set themselves the task of determining what magnitude of length is equal to a second of time, we would only be able to inform them that for us it

> makes no sense to establish any congruence at all between spatial magnitudes and temporal magnitudes. And to the confident assertion that though the task is difficult, surely it could be resolved in principle, we could only respond by appeal to the immediate data of intuition, which teach us just the contrary. The situation is exactly the same when people seek to establish the congruence of two different increments in sensation. Considering the confidence with which we expect uniformity in the basic organization of human intellect (and likely we are justified in general), I do not doubt other people will arrive at the same results by impartial examination. (Niall [1995], p. 291)

The reason why von Kries thought we cannot conceive of increments of sensation being comparable is that different increments of sensation differ qualitatively. The 'uniformity of elements that marks our conceptions of time and space is simply lacking in intensive series of sensations' (Niall [1995], p. 291). Or, in a different translation, 'sensations lack that "kind of sameness that elements have, an equality that is characteristic of our notions of space and time"' (Heidelberger [2004], p. 227).[3]

3. The Quantity Objection: Philosophical Analysis

On the surface, these criticisms—from that of medieval physicists to those of von Kries—resemble each other. To assess their merits, however, we need some distinctions. I will start by distinguishing two criticisms that human measurements typically face and that, though related, we should not consider to be strictly speaking versions of the quantity objection:

(1) The measurement procedure at stake (say, using a life satisfaction scale to measure happiness) does not justify the belief that the resulting measurement results are quantitative. The relationship between someone's actual degrees of happiness and our measurements of them through self-reports might perhaps be logarithmic (versus linear), or we might ignore the shape of this relationship altogether.

(2) The attribute we are trying to measure is not measurable because quantitative measurement is only possible when we can rely on a physical operation of concatenation (say, laying rods end-to-end to measure length).

[3] I highlighted only some of Tannery's and von Kries's criticisms. Another long-lasting criticism they raised is the so-called stimulus error (Boring [1921]; Heidelberger [2004]). Chirimuuta ([2016]) discusses the contemporary relevance of the stimulus-error critique.

The first criticism is commonly launched against human science measurements today. It is popular among methodologists, but also among contemporary philosophers (Wodak [2019]; Ingelström and van der Deijl [2021]). Although they are sometimes run together, the first criticism is distinct from the quantity objection. The quantity objection targets the attribute at stake, not our efforts of measuring it. In contrast, the first criticism targets a particular measurement procedure. It says of a specific procedure—say, using numerical self-reports for quantifying happiness, or using a statistical model such as the Rasch model together with carefully selected test items for quantifying reading comprehension[4]—that it does not deliver quantitative measurement. It does not say that the attribute at stake is not quantitative.

Moreover, some versions of the first criticism presuppose that the attribute is quantitative, so they presuppose the falseness of the quantity objection. For instance, if one claims that the relationship between the degrees of an attribute and our measurement results might be logarithmic instead of linear (Wodak [2019]), then one is thereby assuming that the attribute is quantitative. Unless the attribute is quantitative (and has been already represented quantitatively), it cannot be logarithmically related to something else.[5]

What about the second criticism? Typically, in the right circumstances, concatenation operations enable successful quantification. But the second criticism goes beyond acknowledging concatenation operations as the paradigm of quantitative measurement. It assumes that only with concatenations we can quantitatively measure. The second criticism was popular in Fechner's times, it applies straightforwardly to sensory measurements (we cannot concatenate sensations), and some of Tannery and von Kries's claims may be read as endorsing it. However, nobody defends the second criticism nowadays—it rules out the measurement of temperature, of planets' masses, and so on where there is no concatenation procedure available to us. With the development of non-extensive measurement axiomatizations in the 1960s (Luce and Tukey [1964]), the second criticism lost its appeal entirely. More importantly here, the second criticism

[4] For textbooks on the Rasch model, see (Wilson [2005]; Bond and Fox [2015]); for the evidential value of the Rasch model in establishing quantitative measurement, see (Michell [2008]; McClimans et al. [2017]; Trendler [2019]; Vessonen [2020]).

[5] Unfortunately, Wodak ([2019], p. 30) cites Michell's ([2012a]) argument that psychological attributes are themselves ordinal as an argument that may support the claim that the relationship between happiness and our measurements of it are 'non-linear'. If happiness is itself ordinal, then of course it is not the case that happiness is linearly related to (numerical) happiness measurements. But this is because linearity is not an option, not because happiness could be 'non-linearly' (for example, quadratically, logarithmically) related to measurement results.

demands that a concatenation procedure is available to us. So, it fails to target the attribute itself (versus our access to it). Hence, it is not a version of the quantity objection as understood here.

Having mentioned two common criticisms that (though related) are not versions of the quantity objection, I now distinguish four versions of the quantity objection:

Extensiveness: The attribute is not quantitative because it fails to satisfy the following condition (call it 'extensiveness'): an object's magnitude (say, a sensation's intensity, a rod's length) is determined or 'inherited' by the magnitude of its parts (see Perry [2015]).

Conceivability: The attribute is not quantitative because we cannot conceive the magnitudes of the attribute as being additively related.

Homogeneity: The attribute is not quantitative because it is not homogeneous—its different degrees differ qualitatively among each other—and only homogeneous attributes are quantitative.[6]

Causal Bases: The attribute is not homogeneous (hence non-quantitative) because the causal bases determining the different degrees of the attribute are diverse.

These objections all address the attribute at stake versus our empirical access to it, hence they are all versions of the quantity objection. 3Extensiveness states a metaphysical condition about how objects instantiate the attribute at stake: objects inherit their degrees of attribute X from their parts' degrees of attribute X.[7] Planets inherit their mass from the mass of their parts even if we cannot measure them with a concatenation operation. Hence, unlike the second criticism,

[6] I follow Michell's usage regarding the degree–magnitude distinction, reserving 'magnitude' for quantitative attributes only and using 'degree' neutrally.

[7] On terminology: Perry ([2015]) calls 'additive' what I call extensive. I use 'additivity' as Michell ([1999], pp. 53–54) does: to state how the degrees of quantitative attributes relate to each other (so not to speak of how objects instantiate attributes). Measurement theorists have typically called 'extensive' attributes that we can measure using concatenation operations.

extensiveness does not rule out the measurement of planets' masses. But it does rule out temperature or density being quantitative. (A system in thermal equilibrium has the same temperature as any of its parts, not the sum of its parts' temperatures.) Hence, although historically it has been the most common quantity objection, extensiveness is untenable today—nobody doubts whether density or temperature are quantitative attributes. Therefore, we must not conflate it with other theses.

Recall how Crombie ([1961]) described the medieval argument against temperature's quantitativeness: 'The heats of two bodies brought into contact did not make a greater heat, as the lengths of two bodies [laid out in the right way] made a greater length'. This claim gets close to stating extensiveness, but not quite, since 'make' here is used in a causal sense. Where we saw it clearly stated was in Boring's intended summary of the quantity objection: sensations of higher intensity are not made up of sensations of lesser intensity.

What is less clear, *pace* Boring, is whether von Kries's critique included extensiveness. Some passages, though not the ones I cited above, are suggestive of it. Consider this: 'A loud tone does not conceal within itself this or that many faint tones, in the same sense that a foot contains twelve inches or a minute contains sixty seconds' (Niall [1995], p. 292). The first part of the sentence, about loud and faint tones, arguably expresses extensiveness. But in the second part, von Kries stopped referring to objects that instantiate different degrees of attributes and began talking about degrees themselves (using as names for degrees the standard units). The contrasts he made between loudness on the one hand and length and duration on the other conflated the object-attribute distinction. Let me expand on this, since clarifying this is crucial for not conflating extensiveness with other versions of the objection.

That a foot contains twelve inches, or a minute sixty seconds, are not claims that express extensiveness. They are not claims about how objects instantiate certain attributes (length, duration) satisfying a particular metaphysical-mereological requirement. They are claims about the units used for measuring attributes, hence, claims abstracted from the objects. Moreover, they presuppose that the attributes in question are quantitative. They are analogous to saying that a ten-degrees-increase in temperature on the Kelvin scale—something for which we lack a concise name such as 'foot' or 'minute'—involves ten one-degree increments. These claims presuppose that length, time, and temperature are quantitative, that is, that their degrees are not only orderable but also stand in 'distance relations' (Eddon [2013]) because they are additively related.

Seconds add up, inches add up, temperature degrees add up—this is why seconds, inches and degrees of temperature stand in 'distance relations'.[8] That these attributes' degrees (versus the objects instantiating them) are additively related is not a controversial metaphysical thesis about how objects instantiate attributes. It is just part of the meaning of being a quantitative attribute (Michell [1999], pp. 53–54). Only when the attribute is conflated with the object, an uncontroversial presupposition ('all quantitative attributes satisfy additivity') becomes a controversial metaphysical thesis ('all quantitative attributes are extensive'). As mentioned, the latter is untenable today.

It is clear from the passages I cited above that Tannery's and von Kries's criticisms of Fechner's work are not well captured by Boring's summary of the quantity objection. Even if sometimes they conflate the object-attribute distinction, extensiveness is mostly absent. Conceivability and homogeneity are not, however.

Conceivability doesn't say that the attribute is not additive. As argued, this would be like saying the attribute is not quantitative (hence begging the question). Rather, it invites us to acknowledge that, upon reflection, when we analyse how we think about the attribute, we realize that we cannot imagine it being quantitative, we cannot picture its degrees as adding up (like seconds add up). We might, perhaps out of enthusiasm for putting psychology into physics' quantitative path, start talking about sensations' intensities as quantitative; but conceivability is reminding us how incompatible this idea is with how we actually conceive of sensations. This position is evidently present in von Kries's quotes.

Now, how are we supposed to arrive at this conclusion? Presumably, researchers seriously trying to quantify an attribute (versus merely using measurement procedures that happen to deliver numbers but without any commitment to these numbers informing us about 'distance relations') such as Gustav Fechner are indeed conceiving of the attribute as quantitative. So, despite von Kries's optimism regarding the universality of his intuitions, conceivability on its own is not likely to sway optimists about human quantification. What the critics must offer is an argument for why nobody should conceive of a given attribute as quantitative; not merely report that they do not actually conceive it that way.

What about homogeneity? The intuitive idea here is that an attribute must be qualitatively the same throughout to be quantitative—it must be 'homogeneous' said Tannery, 'uniform' said

[8] For a formal definition of additivity, see (Michell [1999], pp. 53–54).

von Kries (their translators, admittedly). If it is not qualitatively the same throughout, distinct increments of it may be increments of different kinds of things, hence not comparable, hence not additively related.

It is worth repeating the clarification made above. Homogeneity requires, said Tannery, that 'whatever is added, such that something increases, is of the same exact kind as that which was already there: length, surface, and time are dimensions of this kind'. He must be talking about the attribute (say, length), not about concrete objects (rods). The 'something' that increases must be the attribute's magnitude, not the object. After all, laying metal and wooden rods end-to-end is surely not putting things of the same kind together if we are talking about objects, but it is if we abstract and talk about their length magnitudes. Hence homogeneity cannot be a necessary feature of objects, only of attributes. The issue, therefore, is whether the sensations' different intensities (versus sensations themselves) are homogeneous.

That an attribute must be (in some sense to be made precise) 'homogeneous' for it to be quantitative seems uncontroversial—different increments of it must be of the same kind to be comparable. Judgements about homogeneity, however, may not be any more consensus-generating than judgements about quantitativeness. They are, after all, judgements about abstract concepts, not about concrete objects and how they relate. And we have not been given an account of 'homogeneity', despite how much work that term is meant to be doing here. One worry, for instance, is that Tannery (or Heidelberger speaking for Tannery) framed the argument in terms of attributes being either homogeneous or not; but, then, when he gave the example of heat sensation, he admitted that although 'the original sensation (heat) is of an entirely different kind than the final sensation (pain) […] [and that this] also holds for all those [sensations] in between', for these sensations in between this holds 'to a lesser degree'. That is, first he considers homogeneity as a categorical property, then as a graded one. The question inevitably arises: When do attributes lack 'enough homogeneity' to make a difference? Will any small failure of homogeneity do? Why? We are not told.

As we saw, Tannery apparently felt the need to bring in the causal bases claim to further support the intended conclusion. After stressing that sensations of different intensities fail homogeneity—they are of an 'entirely different' kind—Tannery offered as explanation that 'en-

tirely different nerves are involved in those sensations'. Using contemporary jargon: that different causal bases are behind different degrees of an attribute would, thinks Tannery, shows a failure of homogeneity in the attribute.

Attributes can generally have diverse causal bases, however; especially (higher-level) attributes in which multiple realization is common. Some objects are 'fragile' because of their irregular atomic structure, others due to their weak intermolecular bonding (Choi and Fara [2021]). According to some authors, different substances are 'acids'—or behave 'acidically'—for different (microstructural) reasons (Chang [2012]; Hendry [2016]; but see Thyssen [forthcoming]). And different organisms of the same species can have different 'survival fitness' for different internal reasons (longer legs, better sight). Ruling out the quantification of all these (and many other) attributes just because they are multiply realized appears controversial, to say the least.

So, the quantity objection against psychophysical measurement has not been one but several. Some versions are implausible—they smuggle untenable metaphysical requirements. This is the case of extensiveness and arguably of causal bases. Given that extensiveness is untenable, we must not conflate it with other versions of the quantity objection. Alas, this is precisely what Boring did. His rendering of the quantity objection is problematic not only for failing to distinguish different versions of it, but for summarizing the whole objection in terms of an untenable version. Other versions—conceivability and homogeneity—are not ruled out by current scientific practice. But their force admittedly depends on sharing the right judgements about the attribute at stake; judgements about what 'we' can conceive as being quantitative or judge as homogeneous. *Pace* von Kries's deep hopes, judgements seem to differ across researchers. To this extent, these versions of the quantity objection need not sway the optimists.

4. Michell's Quantity Objection

Michell ([2009], [2011], [2012a], [2012b]) has revived the quantity objection. His arguments are worth attending to. They articulate in a fresh way the more plausible version: homogeneity. Moreover, at stake now are not only the intensity of sensations but psychological attributes in general (abilities, traits, attitudes, and so on). Since some of these attributes are used well beyond psychology—for example, in economics, sociology, and medicine—Michell's argument targets much of human science measurement.

Furthermore, recent philosophy of measurement has not engaged much with the quantity objection. It is not as if the quantitativeness of human science attributes is taken for granted across the board. Related worries about in principle measurability sometimes emerge for specific attributes. For example, Daniel Hausman ([2012]) argues that overall health is not a scalar but a multidimensional property, hence it cannot be quantified in the way (say) length is. But putting aside multidimensionality-type arguments for specific attributes, as a general attitude among philosophers of science, the possibility that 'things like attitudes, preference satisfaction, [and so on] are impossible to measure even in principle' is judged to be 'too remote to be worth considering' (Angner [2013], p. 232). Yet this is the very possibility that Michell is giving reasons for.

4.1. The general strategy: Lack of (complete) homogeneity

Michell ([2012a], p. 261) begins with the claim that human science attributes possess 'discernible features incompatible with quantitative structure,' or, more modestly, features that 'make it *not possible to conceive* of' these attributes as quantitative. So, he endorses conceivability. Moreover, just as Mill, Tannery, and von Kries, Michell thinks the key feature behind this is the lack of homogeneity. Hence, he endorses homogeneity. Any quantitative attribute, says Michell ([2012a], p. 262), is homogeneous: different degrees of it are of the same 'kind'. Different degrees of (say) length differ quantitatively—with respect to the amount of length they involve—but not qualitatively.

So far, Michell's argument looks familiar—homogeneity is the mark of quantitativeness, hence, lack of homogeneity leads to lack of conceivability of quantitativeness. Now, Michell is more specific regarding 'homogeneity'. And rightly so, since he is explicitly arguing not only for the negative thesis that psychological attributes are not quantitative, but also for the positive thesis that they are ordinal. And there seems to be a puzzle here. If we can only compare degrees of an attribute when it is homogeneous, how is that ordinal attributes fail homogeneity yet still afford ordinal comparisons?

Michell distinguishes between two kinds of homogeneity: homogeneity among 'degrees' and homogeneity among 'differences between degrees'. Quantitative attributes have both kinds—they have 'complete homogeneity' (my term): both their degrees and their differences between degrees are homogeneous. Consider length. Different degrees of length—say, 1 cm, 2 cm, and

4 cm—are comparable. Differences between degrees also are comparable: we can compare the difference (in length) between 1 cm and 2 cm with the difference between 2 cm and 4 cm. These two differences differ only quantitatively. As Michell ([2012a], p. 264) puts it: 'They are not just magnitudes of the same attribute (namely, length differences); there is no sense in which they differ other than that they are different magnitudes of that same attribute'. Using R. G. Collingwood's terms, Michell ([2012a], p. 262) says quantitative attributes manifest 'a pure difference of degree'.

Ordinal attributes have some homogeneity, says Michell. This is why they afford talk of there being different degrees of the very same attribute. But they have only the first kind of homogeneity (or 'partial homogeneity', my term); they lack homogeneity among differences between degrees. To illustrate, suppose depression severity is itself ordinal. Although *A* is more depressed than *B* who is more than *C*, the idea goes, the difference in depression severity between *B* and *A* is qualitatively different from that between *C* and *B*. What makes *B* more depressed than *C* (say, being more vulnerable to anxiety) differs qualitatively from what makes *A* more depressed than *B* (say, being more prone to suicidal ideas), which makes these differences intrinsically incomparable. Crucially, the claim is not that we ignore how those differences compare with respect to depression severity (say, because our current measuring instruments are strictly ordinal). The claim is that they do not compare. So, faced with the question, 'how is it possible for there being more/less of an attribute without thereby there being much more or less of this very attribute?', Michell ([2012a], p. 262) would respond that ordinal attributes have 'impure differences of degree'. They only have partial homogeneity, not complete homogeneity.

An important aside: One might find talk about 'impure differences of degree' not fully clear. Less charitably, one might think of the distinction between different kinds of homogeneity as an *ad hoc* solution to the puzzle raised above (namely, how is that ordinal attributes fail homogeneity yet still afford ordinal comparisons?). One can agree to some extent. But let me stress that Michell's writings here cited are the only place where, to my mind, one can find recent efforts to address the question of what is for an attribute to be ordinal or, more specifically, how is it possible for there being more of an attribute without thereby there being much more of this very attribute. Metaphysical analyses of quantities can be found in the literature (Eddon [2013]). But there are no metaphysics of ordinality on offer (to my knowledge) illuminating this question. This is important, since we should not take for granted the existence of ordinal attributes.

Historically, many authors have thought that strictly speaking there are no ordinal attributes (Michell [2012b])—any attribute that admits differences of degrees must be quantitative. Rashdall ([1899], p. 369) expressed a common sentiment when writing: 'We certainly say: "This is *more* pleasant than that". The position that the word *more* does not involve the idea of quantity is so startling that I must excuse myself from further discussion of it until it be developed in more detail than has been the case'.

Moreover, while quantitative scales are typically characterized in ontic terms ('the degrees in the Celsius scale are equally distanced'), ordinal scales are commonly described in epistemic terms ('we do not know the distances between the degrees'). To illustrate, this is how Wodak ([2019], pp. 29–30) introduces measurement scales: 'With mental states, as with crowd sizes, it is often easy to know that *A* is greater than *B*, but hard to know the magnitude of the difference between them. We might know that Obama's 2009 Inauguration drew a larger crowd than Trump's 2017 Inauguration did, or that Michelle is happier than Melania, while being ignorant of *how much* larger Obama's crowd was, or *how much* happier Michelle is. In such cases, an *ordinal scale* is appropriate. This is a rank order: Michelle's happiness > Melania's happiness'. Wodak here describes ordinal scales epistemically. He uses our lack of quantitative information to characterize ordinal scales. This epistemic characterization, however, makes sense only if the attribute is quantitative. If the attribute is ordinal, then those 'distances' don't exist, so talk about our ignorance of them is a category mistake.

To be clear, we can formally characterize ordinal variables in non-epistemic terms. Formally, an ordinal variable is one in which transitivity, anti-symmetry, and strong connexity hold (Michell [1990], p. 52).[9] But these formal conditions also hold for quantitative variables. The question that interests us here is not the formal characterization; it is how there can be properties that satisfy these formal conditions yet don't satisfy the further conditions specific to quantitative variables (additivity).

All said, though at first sight implausible, upon reflection one may think that there isn't any (scientifically interesting) ordinal attribute out there. Perhaps there are only two kinds of attributes: quantitative ones, some of which we have not managed to measure quantitatively, and

[9] Anti-symmetry: For any two degrees *a* and *b* of an attribute, if *a* is at least as great as *b* and *b* is at least as great as *a*, then *a*=*b* (equivalently, no two distinct degrees occupy the same place in the ranking). Strong connexity: For any two degrees *a* and *b* of an attribute, either *a* is at least as great as *b* or *b* is at least as great as *a* (equivalently, the ranking is complete).

nominal ones. Further support comes from the fact that common criticisms to the quantitativeness of attributes—say, that the attribute is multidimensional (Hausman [2012]), or that the attribute is just a wrong theoretical characterization of the phenomena of interest, as in Krantz' ([1991]) criticism of utility as a scalar quantity—entail that the attribute is not ordinal either.

The immediate upshot for our discussion is this: those who want to argue that human attributes are not quantitative but merely ordinal should address 'what is to be ordinal?'. Their arguments look harder to accept (all else being equal) if the conclusion ends up being not that human attributes are ordinal (something human science researchers might be able to live with), but that they are just collections of nominal attributes, or that they don't really exist or make sense. Moreover, unless we know what real ordinal attributes look like (versus what quantitative attributes ordinally measured look like), the critics' conclusion that human attributes are merely ordinal will be hard to establish. Specifying what characterizes (intrinsically) ordinal attributes will allow the critics to offer specific reasons for human attributes being ordinal, as Michell does (see below). In this sense, Michell's contribution is significant.

Back to Michell's argument. To conclude that an attribute is ordinal, we must discern 'impure differences of degree'. Though he believes the argument generalizes quite broadly, Michell ([2009], [2012a]) helpfully illustrates it in detail using a scale of 'functional independence' in the elderly. This scale classifies elder people in terms of their lost capacities to perform certain mobility-related tasks. The levels are: (i) climbing stairs, (ii) transferring to bathtub, (iii) bathing, (iv) walking, (v) dressing upper body, (vi) independent toileting, (vii) transferring to bed, (viii) dressing lower body, (ix) mobility without a wheelchair, (x) bladder control, (xi) performing personal grooming, and (xii) bowel control.

Michell uses this scale for several reasons. Functional independence, intuitively, looks like an ordinal attribute, hence it should be easier to illustrate heterogeneity with it. Moreover, because functional independence is not a mental (but a socio-physical) property, we can focus on the ordinality–quantitativeness issue avoiding questions specific to mental attributes (Michell [2009]). Furthermore, the scale has been said to produce data that the Rasch model fits (Embretson [2006]), and it is common among psychometricians to treat the latter as evidence of quantitative measurement (see note 4). Although Michell has challenged that the Rasch model is up to this task before ([2008]), Michell's ([2012a]) point is more fundamental—functional independence is itself ordinal, hence the merits of the Rasch model (or lack thereof) are irrelevant. Michell

([2012a], [2020]) also illustrates his argument, though in less detail, with cognitive abilities. As we will see, the argumentation is similar.

4.2. Applying the strategy

Michell ([2012a], p. 263) explains in which sense he thinks functional independence should be considered an ordinal attribute as follows: 'A person able to complete all of the activities that another person can plus at least one further activity on the list is the more independent of the two. Different degrees of functional independence are, consequently, mutually homogeneous'. According to Michell, then, functional independence is ordinal because (or whenever) peoples' sets of capacities are subsets of each other's. This pattern—peoples' sets of capacities are subsets of each other's—occurs if the capacities the scale mentions are lost in order as one ages, which is something that typically happens (Embretson [2006]; Michell [2012a]).

What about differences between degrees? Michell ([2012a], p. 263) says:

> […] differences between degrees of this attribute are mutually heterogeneous. For example, the difference between being able, on the one hand, and being unable, on the other, to climb stairs, and the difference between being able and being unable to transfer to a tub are differences of qualitatively diverse kinds, and each of these in turn is of another kind to that between being able and being unable to bathe independently, and so on for the other differences between degrees of this attribute. Because of this, the differences between degrees of functional independence do not stand in intrinsic relations of greater than, less than, or equality to one another.

Moreover, continues Michell ([2012a], p. 264):

> Trying to think of differences between degrees of functional independence as thoroughly homogeneous raises the question, what could a decrease in functional independence be other than an inability to do some kind of specific daily activity that one was previously able to do independently? There is no homogeneous stuff, *independence*, adhering in various amounts to each person; there is only the set of distinct capacities to do the range of different daily activities constituting total functional independence, and which, in being lost, successively mark decreasing degrees of that attribute. Thus, functional independence is a merely ordinal attribute.

So, degrees of functional independence are mutually homogeneous, yet differences between degrees are not. This constitutes intrinsic ordinality. Hence, functional independence is itself ordinal.

This argument, Michell claims, is not restricted to functional independence; it holds generally for abilities, traits, and attitudes. Consider cognitive ability ([2012a], p. 265, [2020], p. 316). Tests used to measure mathematical ability (or reading comprehension) include a diversity of items (questions) of varying difficulty:

> Each item of an ability test corresponds to a degree of the relevant ability, namely, that required to pass the item. Were the cognitive resources (knowledge, skills, strategies, etc.) involved in getting any such item correct specified, it would be apparent that as the items increase in difficulty, differences between the required cognitive resources will […] typically […] be heterogeneous. For example, the differences between cognitive resources needed to solve easy and moderately difficult mathematics items will not be the same as the differences between resources needed to solve moderately difficult and very difficult mathematics items. (Michell [2012a], p. 265)

Since items increase in difficulty but also in what kind of cognitive capacities they demand, increments in (say, mathematical) ability as indicated by tests are heterogeneous, constituting intrinsic ordinality.

Before assessing Michell's argument, two qualifications are worth making to avoid misrepresenting his overall position.[10] First, my focus is on Michell's ([2012a]) homogeneity-based quantity objection. This argument, however, is set up within a broader argument that Michell articulates about the plausibility of psychological attributes being quantitative. This broader argument includes other considerations: about the kind of theories that we currently have characterizing psychological attributes—theories which are largely qualitative in character—and considerations against using abduction to defend the quantitativeness of psychological attributes. As said, my focus here is only on his homogeneity-based quantity objection—this objection has been historically prominent and it is the only one that can claim to settle the debate once and for all. (Considerations about the current status of theories cannot establish the impossibility of future quantitative measurement.) Although I criticize Michell's homogeneity-based quantity objection argument below, my conclusion is not at odds with Michell's broader argument.

Second, Michell's intended conclusion can be read in two ways. In the passages I just quoted, Michell takes his argument to establish that human attributes are ordinal. In other passages he's more cautious, concluding that the argument gives a '*prima facie*' reason for human attributes being ordinal (Michell [2011], p. 248) or that this is the most plausible conclusion (Michell

[10] I thank Derek Briggs for useful comments on this and related issues.

[2012a]). His intended conclusion is probably the weaker, even if some passages express the stronger.

4.3. Assessing Michell's argument

In articulating his challenge, Michell characterizes the attributes—functional independence and cognitive abilities—as mere collections of qualitatively dissimilar capacities. For Michell 'there is no homogeneous stuff, [functional] independence […]; there is only *the set of distinct capacities* to do the range of different daily activities *constituting* total functional independence' (emphases added). He does not think of the attribute to be measured as something that, say, causally accounts for these detectable capacities. Rather, functional independence or mathematical ability just are these collections of capacities. Because these capacities differ qualitatively, the attributes are non-quantitative.

I see three problems with this characterization. First, Michell's contrast between functional independence (or cognitive abilities) as a set of diverse capacities versus as homogeneous 'stuff' is problematic: Not all quantities are stuff-like. Length and mass are, but gravitational force isn't.[11] Or consider temperature in early thermometry: if conflated with heat (as was common before Joseph Black's work; see Mach [1986], p. 147–49) and heat is understood as a fluid—say, 'caloric'—then temperature is stuff-like since fluids are. But if temperature is understood as a force—the 'force of heat' (Barnett [1956], p. 331)—or as average kinetic energy, or as the derivative of a body's internal energy with respect to its entropy, it's not stuff-like anymore. More generally, dispositional properties are not stuff-like. If to be stuff-like is to satisfy the condition required in extensiveness, then the contrast that Michell offers is misleading—failure of extensiveness doesn't rule out quantitativeness (something Michell ([1999], pp. 53–54) himself has been keen in clarifying).

Second, *pace* what Michell promised, his characterization leaves functional independence and cognitive abilities not even being ordinal attributes. As Michell states ([1990], p. 52), an ordinal variable satisfies transitivity, anti-symmetry, and strong connexity. Strong connexity (or completeness) fails under Michell's characterization of the attribute. Think that whenever two people lose their functional capacities in different order—that is, whenever the capacities of

[11] Thanks to Miguel Ohnesorge for raising this point.

neither individual are a subset of the capacities of the other—under Michell's characterization of what functional independence itself is there is no fact of the matter of who has more functional independence. It is not as if we ignore who has more functional independence; rather, neither person actually has more than the other nor do they have an equal degree of independence. Same with mathematical ability: students who have the cognitive capacities to correctly answer a fewer number of mathematical questions than their peers do not always have a subset of their peers' capacities. So, *pace* what Michell promised, his characterization entails that the attributes are not even ordinal. (Note that the point is not one of measurement error. Even if the measurement procedure (say, physicians filling in a form, students attempting to answer mathematical questions) is error-free in the sense that the presence or absence of each functional capacity or cognitive capacity is always manifested in the data, we would still have a failure of the subset relationship (therefore of ordinality) as long as elders don't lose their functional capacities and students don't acquire their cognitive capacities in the exact same order.)

Finally, and most importantly for us, Michell's characterization is not one that the subtle optimistic human scientist accepts or must accept. Psychometricians—especially the nuanced optimists—have typically not thought of their attributes as that which is 'directly experienced' (Michell [2012a], p. 264) or immediately captured in measuring instruments like the scale Michell mentions. Rather, theoretically inclined psychometricians have built into their view of measurement the causal complexity of the world, which defies identifying the measurand with what can be straightforwardly read from simple measuring instruments and forces us to understand measurands as theoretical posits.[12]

For contrast, here is a well-known psychometrician characterizing their measurement practice as a strategy to handle causal complexity (what he calls 'Confusion'): 'Confusion is caused by interdependencies. As we look for tomorrow's probabilities in yesterday's lessons, confusing interactions intrude. Our resolution of confusion is to represent the complexity we experience in terms of a few shrewdly invented dimensions … The method we use to control confusion is to enforce our ideas of unidimensionality. We define and measure one invented dimension at a time … *Models which introduce putative causes as separately estimable parameters are our*

[12] Note that physicists are also forced to think of even length or mass as 'theoretical concepts' (versus observational properties) as soon as they go beyond the measurement of middle-size objects in laboratory conditions (Carnap [1966], pp. 102–4).

laws of quantification. These models define measurement, determine what is measurable, decide which data are useful, and expose data which are not' (Wright [1997], p. 38 emphasis added).

This quote illustrates a tradition that is prominent in psychometrics and in other social sciences. Here the measurand is a theoretical property ('an invented dimension')—a parameter in a model that is postulated to causally account for the observable data (Cronbach and Meehl [1955]; Wright [1997]; Borsboom [2005]; Wilson [2005]; Stenner et al. [2013]). From this perspective, the attribute is not constituted by the collection of capacities mentioned in a scale or implicit in a test. Those capacities may be manifestations of the attribute.

Here is how Cronbach and Meehl put it in their seminal paper: 'A construct is some *postulated* attribute of people, *assumed to be reflected* in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct' ([1955], p. 283, emphases added). In Cronbach and Meehl's view, the postulated attribute is theoretical not only in the sense that it is unobservable ('latent'), but also in the sense that its meaning is determined by the role the attribute plays in (postulated and everchanging) law-like relations of theoretical and observable variables (what they call 'nomological networks') (see also Carnap [1966], pp. 102–4). If that is what determines the meaning of the measurand, then, plainly, psychometricians are not conceiving of the measurand as Michell characterizes functional independence or cognitive ability.

Thus described, psychometricians' position is in good company.[13] In contrast with the empiricist- and foundationalist-inspired twentieth-century philosophies of measurement, recent views emphasize the unavoidable role of theoretical assumptions in measurement practices (Chang [2004]; van Fraassen [2008]; Sherry [2011]; Tal [2019]). This emphasis is at odds with thinking of the measurand as that which is straightforwardly read from measuring instruments. Tal's ([2019], p. 873) model-based account of measurement is particularly congenial to the psychometric conception of measurement described above, where measurement is a modelling task heavily informed by theoretical, statistical, and pre-theoretical assumptions (see Borsboom [2005]; Wilson [2005]).

[13] I talk of 'psychometricians' because they are Michell's main target. But treating measurands as theoretical, latent attributes that are causally manifested in measurement indications is prevalent in the human sciences more generally. That said, not everyone takes this approach to measurement. Some areas of psychological measurement are largely theory-avoidant, including those using classical test theory (Borsboom [2005]; McClimans et al. [2017]). The response here articulated, therefore, may not be available to scientists working under classical test theory.

Here is one way of thinking about what these two traditions share. Both (theoretically inclined) psychometricians and theory-laden approaches to philosophy of measurement think of quantification as a long-term bet. Because of the causal complexity of the world, the claim that a given theoretical attribute is quantitative is not a straightforwardly empirically verifiable claim, but a 'working hypothesis' (Sherry [2011]; Bringmann and Eronen [2016]; Tal [2021]). Such a hypothesis 'is introduced tentatively in order to regulate the analysis of data, and then gradually gains (or loses) evidential support by its ability to make data cohere with theories and background knowledge about the measurement process' (Tal [2021], p. 721). That is, the tentative introduction of theoretical quantities (Wright's 'invented dimensions') gets vindicated not so much by surviving one-off tests of quantitative axioms applied to the raw data, but by the gradual coherentist confirmation that putting measurements and theories to use gives rise to.

Taking stock, Michell's argument does not force the (nuanced) optimist to give up their optimism. If optimists conceive the attribute as a theoretical quantity, not as a collection of heterogeneous capacities that we can straightforwardly assess, Michell's argument does not target the measurand as conceived by optimists. Moreover, the optimists' way of conceiving of the attribute is neither unmotivated in general nor indefensible. It shares much with recent theory-laden approaches to quantitative measurement in philosophy. The quantity objection as articulated by Michell, therefore, does not clinch the debate concerning quantification in the human sciences.

5. Final Thoughts: The Ball Is in Scientists' Court

Optimists need not give up their optimism because of the quantity objection. But their defence brings with it serious burdens to discharge; burdens that (many authors claim) are not being discharged by practitioners generally. If correct, this undermines the defence articulated in the previous section, and leaves optimism unmotivated.

That is, postulating quantitative concepts by itself is no response to Michell. Postulation makes sense as part of a research programme—a long-term bet—aimed at the discovery of quantitative relationships, in which postulation is a means for testing. Right now, we lack detailed quantitative theories that can justify treating most human attributes as theoretical quantities suit-

ably coordinated with current measuring procedures.[14] Hence one would expect human researchers to be actively trying to construct and refine quantitative theories about the postulated attributes and coordinating them with concrete measurement procedures.

What should such efforts look like? They need not take just one form, but for illustrative purposes, it's worth considering Richter's programme for the quantification of 'earthquake size'. As argued by Larroulet Philippi and Ohnesorge ([unpublished]), this case shares much more in common with human science quantification efforts than early thermometry does, hence it's more relevant here: First, efforts at quantifying earthquakes were partly driven by obvious social concerns. Second, earthquakes are complex, heterogeneous, and largely inaccessible processes, rendering experimentation of little help for discovering quantitative relationships. Moreover, before Richter, 'earthquake size' was mostly assessed numerically in terms of a synthesis of earthquakes' various effects: on humans (felt by few people, general panic, and so on), on buildings (fall of plaster, partial destruction), and on the earth (trees moving, fissures in the earth's crust). Well-known examples are the Rossi–Forel and the Mercalli scale. Just like (say) depression measurements that combine different symptoms, these scales raise Michell's homogeneity concerns.

Richter's ([1935]) initiated a long-term programme for the quantification of earthquake size. One way of understanding Richter's ([1935]) is as following a strategy for the initial steps of quantification that elsewhere I call the 'causal-residual approach' (Larroulet Philippi [2023]). Here, the postulated quantity (call it X) is provisionally characterized as 'that causal force behind the variations of a specific quantitative observable property (call it Y) when no other factor is affecting Y'. In this approach, increments in X are individuated by whatever is left after removing from the increments in Y what is due to explicitly identified non-X factors. The crucial point is that this way of characterizing the postulated theoretical attribute—quantitatively—enables testing the characterization.

Richter ([1935]) chose as the specific quantitative effect of earthquake size (Y) the maximum amplitude recorded by a particular kind of seismometer. Richter didn't identify 'earthquake size'

[14] On this point Michell is surely correct. He has emphasized throughout his work ([1990], p. 155; [1999], p. 207; [2012a], p. 265) the lack of specificity—of a quantitative character—of psychological theories, alerting that this lack of detail entails that we are not properly able to test quantitative claims about psychological measurands.

with this *Y*. He couldn't—background knowledge entailed that several other factors should affect *Y*. Rather, from the perspective of the causal-residual approach, he (provisionally) characterized earthquake size (calling it 'magnitude') as that force causing variations in amplitude after controlling for the other known or suspected factors. Prominent among these factors were the distance to the epicentre (whose effect on amplitude was modelled), the mechanisms producing earthquakes, the geological structure along which the seismic waves travel, the depth of the earthquake source, and the type of ground below the seismograms. Based on background knowledge, Richter assumed to have roughly controlled for variations on these other factors by limiting his data to earthquakes of (and measurements in) Southern California. This allowed him to test his characterization (or theory) of 'earthquake size'. For example, after correcting for distance, the characterization (plus assumptions) predicts that the same earthquake should be assigned the same magnitude by seismograms in different locations, and the ratio of the recorded amplitudes of any two earthquakes should remain constant across different distances to the epicentre.[15]

These are serious tests. If passed, they suggest we're onto something; if not, they may point towards unsuspected systematic errors. But passing them was only a first step. Why? First, the numbers obtained were highly geographically constrained as well as relatively imprecise; second, such numbers were obtained by assuming away most of what seismologists aimed to discover and that was crucial for improving the robustness and precision of the numbers obtained—namely, whether there are significant differences in the mechanisms producing earthquakes and in the earth's internal structure, the depths of different earthquake sources, how seismic waves behave in different types of ground (crucial for seismic hazard), and so on. Later work by Richter (and Beno Gutenberg) tried to extend the concept of 'magnitude' beyond Southern California, using also other measurement indications based on background theory about wave mechanics. This research led to significant discoveries about the factors abstracted away in Richter's ([1935]), which showed problems with the various magnitude scales, paving the way to a more robust and precise way of quantifying earthquake size (Larroulet Philippi and Ohnesorge [unpublished]; see also Miyake [2017]). All said, progress required committing to a provisional, but specific quantitative characterization of the postulated attribute in terms of its effects, one

[15] These two tests instantiate Chang's ([2004]) 'overdetermination' and Trendler's ([2019]) 'derived measurement' tests, respectively.

that enabled testing, which in turn led to discoveries and successive refinements of the characterization.

There are examples of psychological research programmes specifying the quantitative meaning of the theoretical attribute in ways that enable stringent tests. Notably, some aspects of the Lexile programme for measuring reading ability (Stenner et al. [1983]; Kyngdon [2013]) fit this description. Reading ability here may be thought as that force behind the probability of correctly answering verbal questions once other factors (prominently, item difficulty) are kept fixed. Such understanding of the attribute is standard within Rasch modelling. But the Lexile programme went further, acknowledging the need to quantitatively specify what amounts of 'item difficulty' consist in (using what Stenner et al. [1983] called a 'construct-specification equation').[16] Crucially, this allows (in principle) researchers to test their characterization in the same way as Richter could—testing whether the same student gets assigned the same ability by questions of different difficulty, and whether for any pair of students, the ratio between their odds of correctly answering a question is constant across different degrees of difficulty.[17] (Another notable example is Embretson's research programme, which derives specific characterizations of 'difficulty' for different kinds of tests based on cognitive science modelling. For reading comprehension, see (Gorin and Embretson [2006]).)

These examples—where researchers postulate a specific quantitative characterization of the attribute linked to concrete measurement procedures, thereby enabling stringent tests, which in turn could lead to revisions of the initial characterization and procedures—point in the right direction. Unfortunately, they seem rather exceptional in the human sciences. Well-informed methodologists (besides Michell) complain that what generally goes on in psychometric practice is closer to theory avoidance; unpremeditated usage of off-the-shelf measurement methods and statistical measurement models irrespective of the subject matter and context; and no feedback loops between measurement and theory (McClimans et al. [2017]; McGrane and Maul [2020]; Fried et al. [2022]). To the extent these critics are right, researchers are plainly not working out

[16] Concretely, their quantitative specification combined two dimensions of difficulty: working memory (proxied by average sentence length) and vocabulary (proxied by how rare the words are).

[17] Note that my claim is neither that the Lexile programme has successfully quantified reading ability nor that the specific way in which the programme has been developed (namely, how the quantitative specification of 'item difficulty' was chosen, which tests have actually been performed, the extent to which the test results have led to revisions, and so on) is a model for other research programmes.

the working hypothesis, which undermines the defence articulated in the previous section, and leaves optimism unmotivated. In this sense, the ball is in human scientists' court.

To be clear, nothing said here entails that postulating quantitative concepts and trying to work them out is the only—nor the always preferable—way of conducting research. That the only alternative to quantitative measurement is not doing research is a false dichotomy, as Michell ([2020], p. 312) among many others note. Much valuable research uses qualitative approaches to understand phenomena, prioritizing nuance, context, and particulars at the expense of abstractions and potential generalizability.[18] Quantification efforts need not be the epistemically preferable choice.

There are also ethical or political considerations to keep in mind when evaluating the pursuitworthiness of quantification efforts. After all, representing highly valued skills like reading or mathematical ability as one-dimensional forces lends support to—and arguably has historically been shaped by—hierarchical views of society (Gould [1996]; Anderson [2002]). More generally, choices about what exactly to represent and how specifically to represent it—so-called representational decisions (Harvard and Winsberg [2022])—are choices with potentially moral or political consequences, hence they cannot sidestep moral assessments. For this and related reasons, quantification efforts need not be the morally preferable choice (Larroulet Philippi [2023], pp. 207–11; Ohnesorge [unpublished]). All these, however, are considerations to keep in mind; not arguments that clinch the debate on the possibility of human quantification, as the quantity objection had hoped.

Acknowledgments

Thanks to Anna Alexandrova, Alessandra Basso, Lukas Beck, Derek Briggs, Hasok Chang, Markus Eronen, David Sherry, Jacob Stegenga, Eran Tal, Miguel Ohnesorge, two referees, and participants in the 2022 PSA symposium ('Measuring The Human: New Developments') for helpful comments and suggestions. This paper is based on research funded by the Gates Cambridge Trust.

[18] Coen ([2013]) articulates these trade-offs for the case of early seismology.

Gonville and Caius College

Cambridge, UK

cristianlarroulet@gmail.com

References

Anderson, E. [2002]: 'Situated Knowledge and the Interplay of Value Judgments and Evidence in Scientific Inquiry', in P. Gärdenfors, J. Wolenski and K. Kijania-Placek (*eds*), *In the Scope of Logic, Methodology, and Philosophy of Science*, Dordrecht: Kluwer, pp. 497–517.

Angner, E. [2013]: 'Is It Possible to Measure Happiness?', *European Journal for Philosophy of Science*, **3**, pp. 221–40.

Barnett, M. [1956]: 'The Development of Thermometry and the Temperature Concept', *Osiris*, **12**, pp. 269–341.

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

Bond, T. and Fox, C. [2015]: *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, New York: Routledge.

Bond, T. and Lang, K. [2019]: 'The Sad Truth about Happiness Scales', *Journal of Political Economy*, **127**, pp. 1629–40.

Boring, E. [1921]: 'The Stimulus Error', *American Journal of Psychology*, **32**, pp. 449–71

Borsboom, D. [2005]: *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*, Cambridge: Cambridge University Press.

Bringmann, L. and Eronen, M. [2016]: 'Heating up the Measurement Debate: What Psychologists Can Learn from the History of Physics', *Theory and Psychology*, **26**, pp. 27–43.

Brooks, G. P. and Aalto, S. K. [1981]: 'The Rise and Fall of Moral Algebra: Frances Hutcheson and the Mathematization of Psychology', *Journal of the History of the Behavioral Sciences*, **17**, pp. 343–56

Carnap, R. [1966]: *Philosophical Foundations of Physics*, Basic Books: New York.

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

Chang, H. [2012]: 'Acidity: The Persistence of the Everyday in the Scientific', *Philosophy of Science*, **79**, pp. 690–700

Choi, S. and Fara, M. [2021]: 'Dispositions' in E. N. Zalta (*ed.*), *Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/spr2021/entries/dispositions/>.

Chirimuuta, M. [2016]: 'Why the "Stimulus-Error" Did Not Go Away', *Studies in History and Philosophy of Science Part A*, **56**, pp. 33–42.

Coen, D. [2013]: *The Earthquake Observers: Disaster Science from Lisbon to Richter*, Chicago: University of Chicago Press.

Crombie, A. C. [1961]: 'Quantification in Medieval Physics', *Isis*, **52**, pp. 143–60.

Cronbach, L. and Meehl, P. [1955]: 'Construct Validity in Psychological Tests', *Psychological Bulletin*, **52**, pp. 281–302.

Eddon, M. [2013]: 'Quantitative Properties', *Philosophy Compass*, **8**, pp. 633–45.

Embretson, S. [2006]: 'The Continued Search for Non-arbitrary Metrics in Psychology', *American Psychologist*, **61**, pp. 50–55.

Franz, D. [2022]: '"Are Psychological Attributes Quantitative?" Is Not an Empirical Question: Conceptual Confusions in the Measurement Debate', *Theory and Psychology*, **32**, pp. 131–50.

Fried, E., Flake, J. and Robinaugh, D. [2022]: 'Revisiting the Theoretical and Methodological Foundations of Depression Measurement', *Nature Reviews Psychology*, **1**, pp. 358–68.

Gorin, J. and Embretson, S. [2006]: 'Item Diffficulty Modeling of Paragraph Comprehension Items', *Applied Psychological Measurement*, **30**, pp. 394–411.

Gould, S. J. [1996]: *The Mismeasure of Man*, New York: Norton.

Harvard, S. and Winsberg, E. [2022]: 'The Epistemic Risk in Representation', *Kennedy Institute of Ethics Journal*, **32**, pp. 1–31.

Hausman, D. [2012]: 'Measuring or Valuing Population Health: Some Conceptual Problems', *Public Health Ethics*, **5**, pp. 229–39

Heidelberger, M. [2004]: *Nature from Within: Gustav Theodor Fechner and His Psychophysical Worldview*, Pittsburgh, PA: University of Pittsburgh Press.

Hendry, R. [2016]: 'Natural Kinds in Chemistry', in E. Scerri and G. Fisher (*eds*), *Essays in the Philosophy of Chemistry*, Oxford: Oxford University Press, pp. 253–75.

Ingelström, M. and van der Deijl, W. [2021]: 'Can Happiness Measures Be Calibrated?', *Synthese*, **199**, pp. 5719–46

Krantz, D. [1991]: 'From Indices to Mappings: The Representational Approach to Measurement', in D. Brown and J. E. Smith (*eds*), *Frontiers of Mathematical Psychology: Essays in Honor of Clyde Coombs*, New York: Springer, pp. 1–52.

Kyngdon, A. [2013]: 'Descriptive Theories of Behaviour May Allow for the Scientific Measurement of Psychological Attributes', *Theory and Psychology*, **23**, pp. 227–50.

Larroulet Philippi, C. [2023]: 'On the Challenges of Measurement in the Human Sciences', Ph.D. Thesis, University of Cambridge, available at <doi.org/10.17863/CAM.102194>.

Larroulet Philippi, C. [2025]: 'Against Prohibition (or, When Using Ordinal Scales to Compare Groups Is OK)', *British Journal for the Philosophy of Science*, **76**, available at <doi.org/10.1086/721759>.

Larroulet Philippi, C. and Ohnesorge, M. [unpublished]: 'Case-Selection in Debates about Human Quantification: Lessons from Seismology'.

Luce, R. D. and Tukey, J. W. [1964]: 'Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement', *Journal of Mathematical Psychology*, **1**, pp. 1–27.

Mach, E. [1986]: *Principles of the Theory of Heat*, Dordrecht: Reidel.

McClimans, L., Browne, J. and Cano, S. [2017]: 'Clinical Outcome Measurement: Models, Theory, Psychometrics, and Practice', *Studies in the History and Philosophy of Science*, **65**, p. 67–73.

McGrane, J. and Maul, A. [2020]: 'The Human Sciences, Models, and Metrological Mythology', *Measurement*, **152**, available at <doi.org/10.1016/j.measurement.2019.107346>.

Merbitz, C., Morris, J. and Grip, J. [1989]: 'Ordinal Scales and Foundations of Misinference', *Archives of Physical Medicine and Rehabilitation*, **70**, pp. 308–12.

Michell, J. [1990]: *An Introduction to the Logic of Psychological Measurement*, New Jersey: Erlbaum.

Michell, J. [1997]: 'Quantitative Science and the Definition of Measurement in Psychology', *British Journal of Psychology,* **88**, pp. 355–83.

Michell, J. [1999]: *Measurement in Psychology*, Cambridge: Cambridge University Press.

Michell, J. [2008]: 'Conjoint Measurement and the Rasch Paradox: A Response to Kyngdon', *Theory and Psychology*, **18**, pp. 119–24.

Michell, J. [2009]: 'Invalidity in Validity', in R. W. Lissitz (*ed.*), *The Concept of Validity: Revisions, New Directions, and Applications*, Charlotte, NC: Information Age Publishing.

Michell, J. [2011]: 'Qualitative Research Meets the Ghost of Pythagoras', *Theory and Psychology*, **21**, pp. 241–59.

Michell, J. [2012a]: '"The Constantly Recurring Argument": Inferring Quantity from Order', *Theory and Psychology*, **22**, pp. 255–71.

Michell, J. [2012b]: 'Alfred Binet and the Concept of Heterogeneous Orders', *Frontiers in Psychology*, **3**, available at <doi.org/10.3389/fpsyg.2012.00261>.

Michell, J. [2020]: 'Thorndike's Credo: Metaphysics in Psychometrics', *Theory and Psychology*, **30**, pp. 309–28.

Miyake, T. [2017]: 'Magnitude, Moment, and Measurement: The Seismic Mechanism Controversy and Its Resolution', *Studies in History and Philosophy of Science*, **65**, pp. 112–20.

Niall, K. K. [1995]: 'Conventions of Measurement in Ps§ychophysics: Von Kries on the So-Called Psychophysical Law', *Spatial Vision*, **9**, pp. 275–305.

Ohnesorge, M. [unpublished]: 'We Should Not Align Quantitative Measurements with Stakeholder Values', available at <philsci-archive.pitt.edu/id/eprint/23556>.

Perry, Z. [2015]: 'Properly Extensive Quantities', *Philosophy of Science*, **82**, pp. 833–44.

Rashdall, H. [1899]: 'Can There Be a Sum of Pleasures?', *Mind*, **8**, pp. 357–82.

Richter, C. [1935]: 'An Instrumental Earthquake Magnitude Scale', *Bulletin of the Seismological Society of America*, **25**, pp. 1–32.

Sen, A. [1981]: 'Plural Utility', *Proceedings of the Aristotelian Society*, **81**, pp. 193–215

Sherry, D. [2011]: 'Thermoscopes, Thermometers, and the Foundations of Measurement', *Studies in History and Philosophy of Science*, **42**, pp. 509–24.

Stegenga, J. [2018]: *Medical Nihilism*, Oxford: Oxford University Press.

Stenner, A. J., Smith III, M. and Burdick, D. S. [1983]: 'Toward a Theory of Construct Definition', *Journal of Educational Measurement*, **20**, pp. 305–16.

Stenner, A. J., Fisher, W. P., Stone, M. H. and Burdick, D. S. [2013]: 'Causal Rasch Models', *Frontiers in Psychology*, **4**, available at <doi.org/10.3389/fpsyg.2013.00536>.

Tal, E. [2019]: 'Individuating Quantities', *Philosophical Studies*, **176**, pp. 853–78.

Tal, E. [2021]: 'Two Myths of Representational Measurement', *Perspectives on Science*, **29**, pp. 701–41.

Thyssen, P. [forthcoming]: 'Are Acids Natural Kinds?', *Foundations of Chemistry*, available at <doi.org/10.1007/s10698-023-09485-8>.

Trendler, G. [2009]: 'Measurement Theory, Psychology, and the Revolution That Cannot Happen', *Theory and Psychology*, **19**, pp. 579–99

Trendler, G. [2019]: 'Conjoint Measurement Undone', *Theory and Psychology*, **29**, pp. 100–28.

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

Vessonen, E. [2020]: 'The Complementarity of Psychometrics and the Representational Theory of Measurement', *British Journal for the Philosophy of Science*, **71**, pp. 415–42.

Wilson, M. [2005]: *Constructing Measures: An Item Response Modeling Approach*, Mahwah, NJ: Lawrence Erlbaum.

Wodak, D. [2019]: 'What If Well-Being Measurements Are Non-linear?', *Australasian Journal of Philosophy*, **97**, pp. 29–45.

Wright, B. [1997]: 'A History of Social Science Measurement', *Educational Measurement: Issues and Practice*, **16**, pp. 33–45.