Against Prohibition
(Or, When Using Ordinal Scales to Compare Groups is OK)

Cristian Larroulet Philippi

Abstract

There is a widely held view on measurement inferences, that goes back to Stevens's ([1946]) theory of measurement scales and 'permissible statistics'. This view defends the following prohibition: you should not make inferences from averages taken with ordinal scales (versus quantitative scales: interval or ratio). This prohibition is general—it applies to all ordinal scales—and it is sometimes endorsed without qualification. Adhering to it dramatically limits the research that the social and biomedical sciences can conduct. I provide a Bayesian analysis of this inferential problem, determining when measurements from ordinal scales can be used to confirm hypotheses about relative group averages. The prohibition, I conclude, cannot be upheld, even in a qualified sense. The beliefs needed to make average comparisons are less demanding than those appropriate for quantitative scales. I illustrate with the alleged paradigm ordinal scale, Mohs' scale of mineral hardness, arguing that the literature has mischaracterized it.

1 Introduction

What inferences follow from measurement results? When can we use our measurements to infer, say, that a group of people is on average happier than another? Methodologists answer by pointing to the different kinds of measurement scales. Depending on the information they provide, scales are classified as nominal, ordinal, interval, or ratio. And there are prescriptions regarding (in)appropriate measurement inferences which turn on this classification. This take on the question of measurement inferences goes back to Stevens's ([1946]) theory of scales. Nowadays it is part and parcel of standard quantitative methodology.

Among these prescriptions, the best-known is the prohibition against making inferences from averages taken with ordinal scales (versus quantitative scales). For example, if you measure happiness in an ordinal scale, and one group of people has a higher average on this scale than another group, the prohibition forbids inferring that the first group is happier than the second. Importantly, this prohibition is general—it doesn't target this or that ordinal scale, but any such scale.

Not many scales in the social and medical sciences are widely considered quantitative (interval or ratio). Thus, the prohibition bears tremendously on this research. To illustrate: a large share of these sciences' research concerns causal relations in large populations. This involves taking averages of the measured outcomes among subjects in the treatment and control groups and comparing them to infer causal claims. But precisely this kind of inference is ruled out by the prohibition when the outcomes are not measured with quantitative scales. No wonder that there is frequent discussion about the quantitative status of specific measurements.[1]

The prohibition is patently flouted in social and medical research. To take averages from scales considered ordinal is common practice. Decades ago, economist Richard Easterlin ([1974]) made well-known average happiness comparisons between richer versus poorer individuals and countries. Nowadays all sorts of causal claims are made with such comparisons, from the effectiveness of unconditional cash-transfers (Haushofer and Shapiro [2016]) and housing vouchers (Ludwig *et al.* [2012]), to the impact of children on parents' happiness (Cetre *et al.* [2016]). The same holds for attributes such as life-satisfaction, self-esteem, intelligence, functional independence, depression, among many other.

Because the prohibition is widely endorsed by methodologists across disciplines, condemnation of this allegedly 'illegitimate' practice abounds. In medicine, Merbitz *et al.* ([1989], p. 309) worry about 'the embarrassment to the profession of applying mathematical operations inappropriately' (see also Tennant *et al.* [2004]). In the economics of happiness, Bond and Lang ([2019]) challenge almost all major results in the literature—including Easterlin's famous paradox—because happiness scales aren't quantitative. Most notorious is Joel Michell's analogous criticism of psychological research ([1990], [1997]). Indeed, the tension between ordinary researchers' tools and

[1] As Sherry ([2011], p. 509) puts it, '[p]sychologists debate whether mental attributes can be quantified or whether they admit only qualitative comparisons of more and less. Their disagreement is not merely terminological, for it bears upon the permissibility of various statistical techniques'. I forestall a possible misunderstanding: there is nothing specifically statistic about this issue. Sherry is using Stevens's terminology. But the question of which inferences follow from measurements with different scales is not specific to statistical contexts.

aims—namely, using non-quantitative scales to make inferences about averages—versus methodologists' prescriptions grew strong right after Stevens's articulation of the latter (Lord [1953]). This is 'one of the longest standing debates in behavioural science methodology' said Zumbo and Zimmerman ([1993], p. 390). One that is alive and kicking.

The standard methodological advice about measurement inferences is grounded on Stevens's theory of scales classification and their 'permissible statistics', later axiomatized by the Representational Theory of Measurement (Krantz *et al*. [1971]) (henceforth, RTM). RTM's approach emphasizes the abstract, general mathematical foundations of measurement. Up until recently, this conceptual framework dominated the field. The current 'epistemic turn' in the philosophy of measurement has begun to challenge RTM's dominance in the analysis of measurement practices. We now see philosophical analyses focused more on the background assumptions, and inference patterns behind specific measurement practices (Chang [2004]; Tal [2012]). But this recent focus on concrete measurement practices and inferences has not engaged with the tension between practice and methodology that besieges measurement scales in the human sciences. Here, I aim to contribute to this methodological debate drawing from Bayesian epistemology, the leading contender for a general framework of scientific inferences. Although it has been used to rationalize and/or justify a variety of scientific practices and inferences, as far as I am aware the specific issue of measurement inferences with different kinds of scales has not received a Bayesian treatment.

This paper challenges the general prohibition against making inferences about relative averages unless one has a quantitative scale. I show that under some circumstances, such comparisons can be made with non-quantitative scales. I begin my critique in section 2, showing that the standard argument given for the general prohibition cannot justify it, at least in its commonly endorsed, unqualified form. In section 3, I challenge the qualified form of the prohibition. I use a Bayesian analysis to determine when measurements from non-quantitative scales can be used to confirm hypotheses about relative group averages. The pivotal issue ends up being how different in magnitude specific intervals of the scale are. We don't need to believe all intervals to be equal (as in the case of quantitative scales) to confirm hypotheses about relative averages. Section 4 addresses objections regarding the applicability of this analysis. I show that some of these objections mischaracterize scales widely considered ordinal. I illustrate this point with the alleged paradigm ordinal scale, Mohs' scale of mineral hardness.[2]

A terminological note. Standard measurement methodology, as seen both in methodology textbooks and in everyday methodological disputes among practitioners, doesn't draw distinctions within non-quantitative scales. It considers all non-quantitative scales to fall under the class ordinal (*locus classicus*: Stevens [1946], Table 1).[3] Accordingly, the prohibition against average comparisons with ordinal scales in practice amounts to a prohibition against average comparisons with non-quantitative scales, which is the prohibition I'm challenging. Now, we'll see along the way that it is possible to draw some distinctions within non-quantitative scales, thereby reducing the extension of the term 'ordinal scales'. At that point, we'll see that a narrower version of

[2] As one referee rightly noted, some statisticians also contest the prohibition (Abelson and Tukey [1959]; Labovitz [1967]; Velleman and Wilkinson [1993]; Zumbo and Zimmerman [1993]). Many criticize the standard scale classification and prescriptions in ways congenial to my argument (see subsection 4.1). But their specific arguments against the prohibition are distinct from mine: they are less principled and more results-oriented. Most reject the prohibition with the claim (typically justified with statistical simulations or sensitivity analyses) that assuming one particular cardinalization (say, linearity) of the measurement outcomes doesn't systematically change the statistical results. This argument is orthogonal to mine; I don't address it here.

[3] Putting nominal scales to the side, of course.

the prohibition can be salvaged. Crucially, this version cannot vindicate the way the prohibition is actually deployed in methodological disputes.

2 The Argument for the Prohibition[4]

2.1 Kinds of scales

Ordinal scales, methodology textbooks tell, provide rank orders. For instance, in an attitudinal question, people's attitudes may be classified in {5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree}. Quantitative measurement begins with interval scales. Here the intervals—the differences between subsequent levels of the scale—are equal in magnitude. So, the difference in temperature between 2ºC and 3ºC equals the difference between 4ºC and 5ºC. This equality among intervals distinguishes interval scales from ordinal ones—unlike the case of temperature, we don't know if the distance between 'strongly agree' and 'agree' equals that between 'agree' and 'neutral'. Finally, we distinguish ratio from interval scales because ratio scales have a non-arbitrary zero.

Stevens (and later RTM) formalized these scale types by reference to the 'uniqueness' of their numerical assignment. This uniqueness is, in turn, defined by the transformations to the numerical assignments that preserve the information the scale gives. RTM calls these transformations 'permissible' or 'admissible' (Krantz *et al.* [1971]; Roberts [1985]). For ordinal scales, any order-preserving transformation is admissible. This expresses formally the intuitive idea that ordinal scales inform only about the relative order of elements (Stevens [1946]). Thus, any order-preserving transformation gives us the same information we already had. A standard RTM textbook (Roberts [1985], p. 65) illustrates ordinal scales thus:

[4] This section draws from (Larroulet Philippi [2021b]).

> Some scales are unique only up to order. For example, the scale of air quality being used in a number of cities is such a scale. It assigns a number 1 to unhealthy air, 2 to unsatisfactory air, 3 to acceptable air, 4 to good air, and 5 to excellent air. We could just as well use the numbers 1, 7, 8, 15, 23, or the numbers 1.2, 6.5, 8.7, 205.6, 750, or any numbers that preserve the order.

The idea here is that this air quality scale is not informative about whether the intervals of the scale are of the same magnitude or not. As Merbitz *et al*. ([1989], pp. 308–9) put it, '[a]n ordinal scale, regardless of how it was developed, is defined as ordinal because the distance from [level to level] is not known. Thus, there is no real basis for choosing one number progression instead of another'. In sum, though the usage of {1, 2, 3, 4, 5} in the air quality scale might suggest equality between intervals, from the perspective of standard methodology this is incorrect. Another example mentioned by RTM researchers is the Beaufort Wind Force Scale (Suppes and Zinnes [1963]; Baccelli [2020]), which goes from 0=calm to 12=torment. But the paradigm example is Mohs' scale for minerals' hardness—almost every author illustrates ordinal scales with it, regardless of their affiliation to RTM (Stevens [1946]; Suppes and Zinnes [1963]; Roberts [1985]; Blackburn [1996], Sherry [2011]; Michell [2012]; Tal [2017]; Baccelli [2020]).

For interval scales, any positive linear transformation (that is, a transformation from *x* to y that satisfies: $y = \mathrm{a} + \mathrm{bx}, \mathrm{b} > 0$) is admissible. Any such transformation may change the magnitude assigned to 0 (if $\mathrm{a} \neq 0$) and the value of the intervals (if $\mathrm{b} \neq 1$), but not the intervals' equality. For example, you can translate temperature measurements from Fahrenheit to Celsius using the transformation a=-160/9, b=5/9. As well as having equal intervals, ratio scales have a natural (non-arbitrary) 0. Thus, only positive similarity transformations ($y = \mathrm{bx}, \mathrm{b} > 0$) are admissible.

2.2 The prohibition's justification

Sometimes the justification for the prohibition is just that taking averages from ordinal scales 'does not make sense' (which, as Suppes and Zinnes commented, amounts to 'a phrase often used when no other argument is apparent' [1963], p. 3]). The strongest argument for the prohibition, however, turns on two premises (see Michell [1990], pp. 40-46 for a clear presentation): the set of admissible transformations that define ordinal scales and the following invariance principle. 'When inferring claims from measurement results, only the conclusions that remain true under all admissible transformations are validly inferred'. This principle reflects the fact that scales are defined by what is common across their admissible transformations—all admissible transformations of a scale represent the phenomenon equally well. So, to infer a conclusion when derived only under some specific transformation would be premature: we need to verify whether it is derivable under every admissible transformation. If not, the conclusion doesn't follow. Consider this: If we could infer a conclusion that is not invariant, we could arrive at contradictory conclusions starting from the same measurement results.[5]

In the case of ordinal scales, the set of admissible transformations are the order-preserving ones. Take Mohs' scale. Its assignment of numbers to minerals involves the following procedure: if mineral a scratches mineral b, then a is harder than b. This scale assigns numbers from 1 to 10 in increasing levels of hardness to specific minerals. Here Sherry ([2011], p. 514; author's emphasis) gives the argument for an inference prescription (other than our prohibition) regarding Mohs' scale:

[5] Stevens conceived this issue as one of 'permissible' mathematical operations. RTM saw here a problem of 'meaningfulness': if a claim is not invariant to admissible transformations, it is not empirically 'meaningful' (Krantz *et al*. [1971]; Roberts [1985]). As Michell ([1990], pp. 40-46) argues, the issue is better understood as one of valid inferences (versus permissibility or meaningfulness).

> The arbitrariness of Mohs' scale lies in the freedom available to Mohs in assigning numerals to different levels of hardness. The only restriction on his assignments is that the numeral assigned to the harder mineral come *after* the numeral assigned to the softer mineral, given some procedure for determining which of the two minerals is harder than the other. That is, any monotonic transformation of Mohs' scale would have served Mohs' purpose—viz., ranking—equally well. Relative to Mohs' scale the difference between the hardness of diamond (10) and the hardness of quartz (7) is 3, and likewise the difference between the hardness of quartz and the hardness of fluorite (4) is 3. But one may not infer that quartz is harder than fluorite by an amount equal to that by which diamond is harder than quartz *unless* any monotonic transformation of Mohs' scale would dictate the same inference. Obviously, that condition fails to hold; the transformation f: $n \rightarrow 2^n$, for instance, would not yield an equality.

Focusing now on the prohibition: we should not infer that set of objects A is on average harder than set B from the fact that their hardness levels in Mohs' scale are A={3,3,4} and B={1,3,5}. For if we apply the permissible transformation Sherry used ($y = 2^x$), as well as many others, A's average becomes lower than B's. In sum, a conclusion such as 'set A is on average harder than set B' follows only if it is derivable independently of the order-preserving transformation used.

2.3 Qualifying the prohibition

An unqualified prohibition against making inferences with averages from ordinal scales is commonly endorsed. Joel Michell ([1990], p. 46), to name a prominent methodologist, concludes his presentation of the very argument just given with the following claim: 'So Stevens' prohibition against calculating means of ordinal data does have something to recommend it: ordinal data alone entails nothing about means, so following his prescription excludes no valid consequences of one's data'.[6] Merbitz *et al.* ([1989], p. 309), expressing their frustration with medical research, also state the prohibition in an unqualified way: 'Often […] ordinal scores are manipulated mathematically […] or statistically […] The results of such operations are not logically valid'.

These quotes exemplify the unqualified prohibition, which states: inferences about group averages with ordinal scales are never valid. The argument provided above (subsection 2.2), however, targeted only the inferences that don't hold after permissible transformations. Thus, this unqualified prohibition is not justified by the argument—it overlooks the argument's invariance clause. This is no hair-splitting: the invariance clause does hold sometimes regarding inferences about relative averages, leaving them unchallenged by the argument.

When are conclusions about relative averages from ordinal scales invariant? The concept of stochastic dominance helps to state this condition precisely. Imagine there are two groups of people (A and B) measured by an ordinal scale of happiness that goes from 1 to 3 ('Not very happy', 'Fairly happy', and 'Very happy'), and I want to compare the groups' average happiness so as to conclude which is happier. Let G_A and G_B be the cumulative distributions of each group: $G_A(x)$ is the fraction of people in group A that are as happy or less happy than x, and similar for G_B. A well-known result in statistics and economics says: The computed average of A will be higher than that of B for any order-preserving transformation iff $G_A(x) \leq G_B(x)$ for all x and with a strict inequality over some values of x. This condition is called First Order Stochastic Dominance (Hadar and Russell [1969]) (henceforth, FOSD). Figure 1 shows an example where FOSD holds. Note that G_A and G_B don't cut across: G_A lies on or below G_B everywhere and strictly below over at least one interval, which is what FOSD requires.

[6] In fact, Stevens ([1946]) was more ambivalent than Michell about the prohibition.

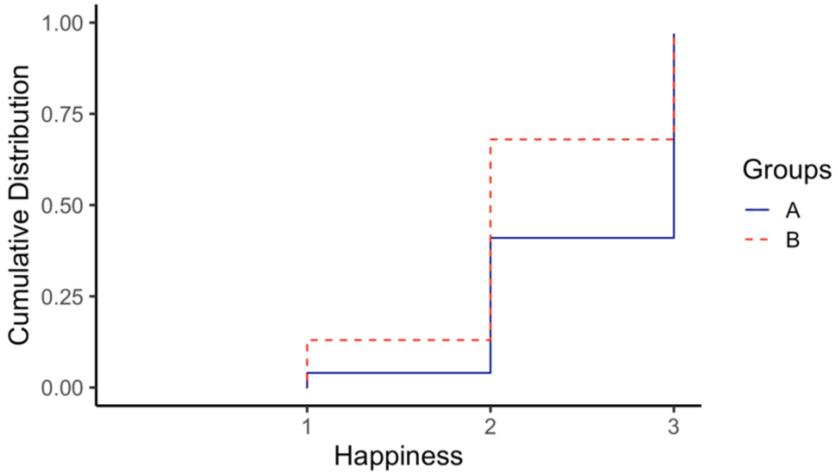

Figure 1. Cumulative distributions of groups A and B. FOSD holds.

The data in Figure 1 comes from Easterlin's ([1974], Table 2) happiness comparisons by income groups in the US. A is the highest income the group, B is the lowest. As said, this comparison satisfies FOSD.[7] So, no matter what order-preserving numbers we assign to the categories—e.g., {1, 2, 3}, {1, 3, 100}, etc.— A's average happiness will always come out as higher than B's. The conclusion that group A is on average happier than group B, then, is invariant to permissible transformations. Thus, we cannot say this inference is 'not logically valid'. Pace Michel, prohibiting inferences from averages of ordinal scales does exclude valid consequences of the data in some cases.

[7] The probability of being 'not very happy' (1) is 4% for group A (versus 13% for B), of being 'fairly happy' (2) or less is 41% (versus 68%), and of being 'very happy' (3) or less is 97% (versus 97%). The frequencies don't add up to 100 because of the no-answer rate (3% for both groups).

Taking stock, the traditional view on ordinal scales provides *prima facie* grounds for questioning measurement inferences from ordinal scales. Because we don't know that intervals are equal, inferences from averages with ordinal scales lack the epistemic security that quantitative scales provide. To this extent, the traditional view vindicates methodologists' complaints mentioned in the introduction. However, as I have just shown, the argument here considered does not justify an unqualified prohibition against making deductive inferences with averages from ordinal scales. When the differences among groups are strong enough (so that FOSD holds), that argument does not block inferences about relative average comparisons.

FOSD is a demanding condition—many group comparisons of interest don't satisfy it. Are we forbidden to make these average comparisons, as the qualified (but still general) prohibition states? Fortunately not always, or so I will argue. To challenge the prohibition, however, we need a subtler epistemic framework; one that allows for belief states between knowing that all intervals of a scale are equal and knowing nothing about how they compare.

3 Challenging the Prohibition

I start with a concrete, simple example. A formal treatment of group comparisons in general then follows.

3.1 An example

Researcher M proposes to compare the hardness of two sets of minerals (A and B, each containing four minerals), so as to (dis)confirm hypothesis H: 'A is on average harder than B'. Using Mohs' scale, M obtains the following measurements A={1,2,3,8} and B={2,2,3,3}. Had M obtained those measurements with a quantitative scale, she would have inferred H, since A's average hardness (3.5) is larger than B's (2.5). But Mohs' scale is considered the paradigm ordinal scale. So, the standard view on measurement tells this data cannot (dis)confirm H. Moreover, Figure 2 shows the cumulative distributions of both sets of minerals (G_A and G_B): FOSD does not hold, the distributions cut across. Thus, nothing can be said about H. Or so the qualified prohibition says.

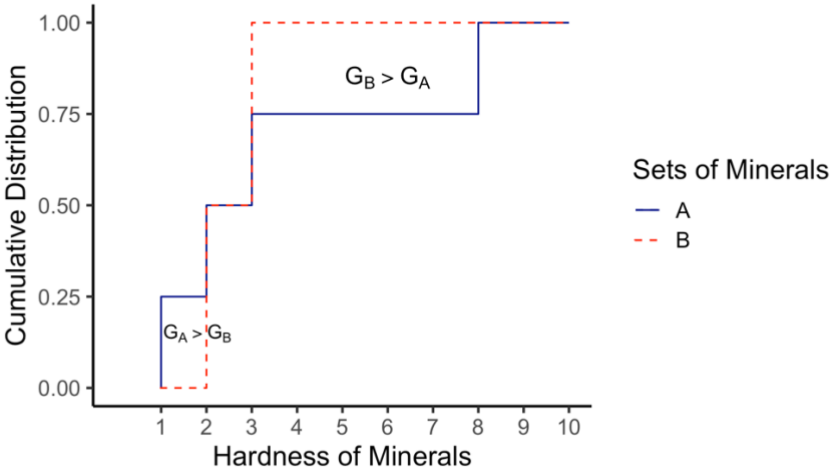

Figure 2. Cumulative distribution of sets A and B. FOSD doesn't hold.

We may ask: What does M have to rationally believe about the scale so as to be justified in using this data to (dis)confirm hypothesis H? Under the standard view, M must believe her scale to be quantitative. But this is not strictly needed. Let us start by focusing on the areas in Figure 2 where the cumulative distributions don't overlap. FOSD holds when there is no stretch along the horizontal axis where G_A is over G_B (G_A>G_B) and there is some stretch where G_A is under G_B (G_B>G_A). The former fails in our example. However, FOSD is a special case of a more general rule, which we may call 'General-Rule-R' (or 'R' for short): the total area where G_A is over G_B is smaller than the total area where G_A is under G_B. The sizes of these areas depend both on the relative sizes of the intervals and on the proportion of elements in the relevant categories (width times height).[8] It is General-Rule-R that tells us whether hypothesis H holds.

[8] Figure 2 depicts the intervals as being all equal, but this of course is not being assumed.

Thus, for inferring H, we don't need to be certain that all intervals are of equal size. That is, we don't need a quantitative scale. At most we need to know the relative sizes of the specific intervals where the groups differ.

To see this, note that when comparing A={1,2,3,8} to B={2,2,3,3}, we can safely remove the two elements in the middle from the comparison, since their contribution to each group's average hardness is the same. We now only need to compare {1,8} to {2,3}. That is, A is on average harder than B as long as {1} and {8} taken together are harder than {2} and {3} taken together. In terms of intervals' differences: A is harder on average than B iff the degree to which {2} is harder than {1} is less than the degree to which {8} is harder than {3}. I will call this condition 'Interval-Difference-K' (or 'K' for short): that relation between intervals' differences that is necessary and sufficient in a given case for hypothesis H to hold. Inspecting Figure 2, we see that Interval-Difference-K is precisely what the comparison of areas involved in General-Rule-R requires: the area where G_A>G_B is smaller than the area where G_B>G_A iff the difference in width (degree of hardness) of {2} vs. {1} is smaller than the difference of {8} vs. {3}. (We say 'width' and not 'area' because the height remains the same in this example.) So, K iff R.

Now we see clearly what makes FOSD special. When it holds, General-Rule-R (recall, the rule about areas' relative sizes) is satisfied no matter what the relative magnitudes of the intervals are. This is so because there is no comparison of areas to be made: there is an area where G_B>G_A but there is no area where G_A>G_B. When FOSD holds, after netting out the objects that have the same values of hardness (in the way we did in the previous paragraph), what remain are objects in set A that are all harder than every object remaining in set B. That is why G_A is never over G_B. In contrast, in our example, set A has one object, {1}, with less hardness than those that remain in group B ({2,3}). Because for all we know, it is possible that Mohs' scale has heterogenous enough intervals, such that the degree to which {2} is harder than {1} surpasses the degree to which $\{8\}$ is harder than $\{3\}$, it is possible that H doesn't hold. It is also straightforward to see what makes quantitative scales special. This last possibility—of a single interval being bigger than five other intervals combined—is ruled out. We know all intervals are equal.

By making clear what's special about FOSD, however, we also see why we may not need it. To make average comparisons, we need FOSD to hold only if we cannot rule out any possible difference between intervals. As long as we can rule out some possibilities, we may not need FOSD for inferring the hypothesis of interest (H). In our example, General-Rule-R holds if the magnitude of hardness between $\{8\}$ and $\{3\}$ is larger than that between $\{2\}$ and $\{1\}$. Thus, what M has to rationally believe in order to deduce H is merely that the size of the 1-2 interval is less than five times larger the average size of the intervals 3-4, 4-5, 5-6, 6-7, 7-8. This is a very different—and much weaker claim—than that each interval of the scale is of equal size. Perhaps the evidence does not warrant believing this weaker claim. Perhaps it does. But the point is that neither the belief of having a quantitative scale nor of FOSD holding are needed, if the researcher rationally believes that the differences in intervals are such that General-Rule-R holds.

3.2 Formal Analysis

Working out a formal representation of this inference problem helps drawing the general lessons. We can model the difficulty involved in working with non-quantitative scales as a case of 'unreliable evidence' (Howson and Franklin [1994]). In our context, the source of unreliability comes from the potential heterogeneity of the specific intervals that matter for the average comparison. This is captured by Interval-Difference-K (recall, that relation between intervals' differences that is necessary and sufficient in a given case for hypothesis H to hold).

Before going forward, however, let me acknowledge the somewhat idealized nature of the discussion that follows. My focus lies on the quantitative versus ordinal scales issue. Thus, the only source of uncertainty modelled below is that about the intervals' differences. There are other sources of uncertainty (on measurement error, see Kyburg [1992]). These other sources, which no doubt should affect the firmness of our beliefs—such that, for example, we plausibly never assign credence 1 to contingent propositions—are assumed away below. But those uncertainties apply equally to both kinds of scales, and don't play any role in the prohibition's justification. So we don't need to consider them. (Thus, when I say below 'we should assign credence 1' to some proposition, this is meant as a claim relative to a background where the only uncertainty is that of intervals' differences.)

Imagine again researcher M using Mohs' scale for (dis)confirming a relative average hypothesis H (such as 'set A is harder than set B'). She measures two sets of minerals, gathers the data, and then deduces the Interval-Difference-K of this specific comparison (that is, the condition about intervals' differences that is necessary and sufficient for H to hold given the specific data observed). In the example above, K: 'the 3-8 intervals combined are larger than the 1-2 interval'. In different group comparisons, K would be a different condition. But in every group comparison, there is a K such that H is true iff the specific K of the case holds. Hence, when M observes her measurements, she learns a biconditional, $E: H \leftrightarrow K$.

We can assume that, prior to the measurement, H and K are probabilistically independent. The former is about the relative average hardness of two specific sets of minerals, the latter about the relative differences between some specific intervals of a measurement scale. The likelihood is $P(E \mid H) = P(H \leftrightarrow K \mid H) = P(K)$.[9] And thus the researcher's posterior is:

$$P(H \mid E) = P(H) * P(K)/P(E).$$

This result is intuitive: H's posterior probability depends proportionally on how likely K is to be true. Developing $P(E)$ via the law of total probability, we have that

$$P(E) = P(E \mid H\&K) * P(H\&K) + P(E \mid {\sim}H\&K) * P({\sim}H\&K) +$$
$$P(E \mid H\&{\sim}K) * P(H\&{\sim}K) + P(E \mid {\sim}H\&{\sim}K) * P({\sim}H\&{\sim}K).$$

Because the two middle terms are 0, and both $P(E \mid H\&K)$ and $P(E \mid {\sim}H\&{\sim}K)$ equal 1,

$$P(E) = P(H\&K) + P({\sim}H\&{\sim}K) = P(H) * P(K) + P({\sim}H)\ P({\sim}K).$$

The posterior is then a weighted average of M's confidence about K:

$$P(H \mid E) = P(H) * P(K)/[P(H) * P(K) + P({\sim}H)\ P({\sim}K)].$$

The higher the priors of H and K, the higher the posterior. If M is certain that K holds, then H's posterior is 1, just like in the quantitative scale case. Moreover, it is straightforward to see that there is confirmation ($P(H \mid E) > P(H)$) whenever

[9] The steps are as follows. From the definition of the biconditional, we have:

(1) $P(H \leftrightarrow K|H) = P((H\&K) \vee ({\sim}H\&{\sim}K)|H)$.

Because $H\&K$ and ${\sim}H\&{\sim}K$ are mutually inconsistent, we have:

(2) $P(H \leftrightarrow K|H) = P(H\&K|H) + P({\sim}H\&{\sim}K|H)$.

Because the second term is 0, and H and K are probabilistically independent:

(3) $P(H \leftrightarrow K|H) = P(K|H) = P(K)$.

$P(K)$>0.5.[10] This, again, is rather intuitive. As long as the researcher is more confident than not that K, the evidence confirms (in a graded sense) H.

Perhaps having an evidence-based prior on H is not always possible. Are we doomed? Likelihoodism would suggest focusing only on whether the evidence favors H over its competitor, which is ~H here. The evidence does so whenever $P(K)$>0.5, so we arrive at the same result. In this sense, the "problem of priors" is no problem for our argument, and our lesson holds beyond Bayesianism.

3.3 Discussion

From a Bayesian perspective, hypothesis H is confirmed (in a graded sense) whenever $Pr(H|E) > P(H)$. One case where confirmation happens is when the researcher believes that the intervals are such that Interval-Difference-K holds. This indeed occur when the researcher believes her scale is quantitative. But our main lesson is that this latter belief is not needed for rationally believing that K holds. Beliefs less demanding than equality among all intervals— which may be plausible in some cases, not in others —can be sufficient for believing K. Hence, the general prohibition should be rejected: by requiring quantitative scales for average comparisons, the prohibition requires more than what is needed for deducing H from measurement results.

Moreover, the argument for the prohibition tacitly assumes that measurement inferences are restricted to deductive inferences. Our Bayesian framework helps us see this is too restrictive—we don't need certainty about K to make evidence bear on hypothesis H. Researchers need to be more confident than not about K for confirming H. This is a second sense in which the prohibition should be rejected, since it requires more than is needed for making measurement results bear on H. Thus, our major lesson is that a general prohibition, even a qualified one, is not justified. Inferences with averages from

[10] By replacing $P(\sim H) = 1 - P(H)$ and $P(\sim K) = 1 - P(K)$ we arrive at this result.

non-quantitative scales may be valid, and the analysis presented here shows exactly when this is so.

A further lesson to draw is that the epistemology that grounds the general, blanket prohibition is too blunt, and it contrasts substantially with the approach to measurement inferences here considered. To see this, think that not all comparisons of average hardness in ordinal scales where FOSD doesn't hold are on a par. If in our example, the fourth element of group A had a hardness of 10 (instead of 8), we would know that A is in a better position to have a higher average of hardness than B (after all, now A is strictly harder than before using the FOSD criterion.). The converse is true if the fourth element had a hardness of 4. But these variations are not captured in an epistemology that says 'we can only (dis)confirm H if FOSD holds, or if we have a quantitative scale.' Such epistemology cannot distinguish many cases of stronger versus weaker evidence.

To explore this further, let's try to translate the prohibition to the Bayesian framework here entertained. One way is this: 'anytime a researcher faces an example like that of M, her prior on Interval-Difference-K is such that there is no (dis)confirmation.' In this way, it might seem that endorsing the prohibition doesn't go against our approach. How could this strategy work? The result of 'no (dis)confirmation, always' happens if researchers' priors are $P(K)$=0.5 in all cases. If researchers are never more confident than not in K, no H is (dis)confirmed.

The problem: this tactic violates coherence. Imagine M obtaining measurements for sets of minerals A=$\{1,2,3,8\}$ and B=$\{2,2,3,3\}$, just as before, but now also for C=$\{1,2,3,7\}$. To be unable to (dis)confirm the hypothesis H_{AB} that A is on average harder than B, as we saw, M has to believe that $P(K_{AB}) = 0.5$, where K_{AB}: 'the 3-8 intervals combined are larger than the 1-2 interval'. Now, to be unable to (dis)confirm the hypothesis H_{CB} that C is on average harder than B, M has to believe that $P(K_{CB}) = 0.5$, where K_{CB}: 'the 3-7 intervals combined are larger than the 1-2 interval'. But $P(K_{AB}) = 0.5 = P(K_{CB})$ is incoherent. In Mohs' scale, every interval is believed to have positive value, so does the

7-8 interval. Therefore, M cannot be equally on the fence regarding conditions K_{AB} and K_{CB}. Her credences must be $P(K_{AB}) > P(K_{CB})$. Thus M cannot (have beliefs that) uphold the prohibition and satisfy coherence. This much is the prohibition at odds with our Bayesian framework.

*

Our analysis shows when average comparisons are justified. If we plot the cumulative distribution of both groups, General-Rule-R tells us which areas to compare (and thus, which beliefs about intervals are needed, which is what condition Interval-Difference-K expresses). Admittedly, I illustrated this logic with a somewhat simple case, where we needed to compare only two intervals (or combined adjacent intervals). How common are such cases in everyday research practice, and how different are they from more complex ones?[11]

Our imagined example can be considered simple in the sense that the comparison involves an area composed of one interval and an area composed of a combination of adjacent intervals. More complex cases involve comparing areas composed of non-adjacent intervals. The complexity of comparisons increases as the number of times the cumulative distributions of the groups under comparison cut across each other. Thus, complex cases arise only when scales have many levels (otherwise the distributions cannot cut across several times) and cumulative distributions are very heterogeneous (so that they actually cut across). In our example, the scale does have many levels (ten), but the distributions cut across only once.

[11] Thanks to a referee for this question. By calling this case ‘simple’ I’m not suggesting that all researchers will always be in a position to make such comparisons—if they lack substantial understanding of the attribute being measured and the measuring instrument being used, they just may not know how the intervals compare.

Simple cases are widespread: First, it is not uncommon for cumulative distributions to cut across only one or two times, even in scales with many levels. To give one example, the recent *Handbook for Wellbeing Policy-Making* (Frijters and Krekel [2021], Figure 2.3) motivates the usage of the life satisfaction scale (which has eleven levels) for policy evaluation with a comparison between the UK, Denmark, and France. If you plot these cumulative distributions (available upon request), you see the distributions either don't cut across or cut only once.[12]

Second, when scales have few levels (three or four), the comparisons are necessarily simple. Indeed, all comparisons with three levels are simpler than our example since there are only two intervals in play. In scales with four levels, the comparison can be at most as complicated as something like this: '*H* holds iff the 2-3 interval is not (say) four times bigger than the 1-2 and 3-4 intervals combined.' Crucially, scales with few levels are quite common in social psychology, political science (Krosnick and Presser [2010]), and development economics (examples: Easterlin [1974]; Ludwig *et al*. [2012]; Acemoglu *et al*. [2020]).

Of course, more complicated cases than our imagined example also exist. Now, even when the comparison is a complex one (because the distributions cut across several times), the inferential logic remains the same: we need to compare the two (combined) areas as indicated by General-Rule-R. It is only that more non-adjacent intervals need to be combined to compare such areas. All said, the kind of simple comparisons with which we illustrated the solution do generalize.[13]

[12] I thank Kelsey O'Connor for sharing this data.

[13] A thorny issue to consider, as suggested by one referee, is the possibility that the sizes of the intervals vary across individuals. This is a distinct problem than that of ordinality in itself; it is about interpersonal comparability. But it bears on the average comparisons

4 Objections

4.1 Can we have such beliefs?

An objector might grant what I claimed about the validity of inferences, namely that if researcher M holds rational beliefs about intervals' differences she may deductively infer (or probabilistically confirm) hypothesis *H*. But the objector may dispute the antecedent of this claim. Nobody can have such beliefs, because that's precisely what it means to have an ordinal (versus quantitative) scale: we just don't know how the different intervals compare among each other.

We should distinguish two versions of this objection. First, the more ambitious claim: 'Researcher M cannot hold rational beliefs about the intervals' differences of ordinal scales. Thus, M cannot infer hypothesis *H* from ordinal scale measurements (unless FOSD holds). Therefore, the general (qualified) prohibition survives'.

we are entertaining. As the literature testifies, this variability is likely present (in various degrees) in self-reporting scales, such as happiness or life satisfaction scales (though not necessarily for all other kinds of scales). And it is probably present to a larger extent when the comparisons intended are cross-cultural, since people's understandings of the categories might differ cross-culturally. Note, however, that this issue need not affect substantially all average comparisons. First, the extent to which this variability is present is likely to depend on whether the comparison is made within culturally homogeneous groups. (This supports judging the validity of measurements context-dependently, as argued in Larroulet Philippi [2021a]) Moreover, strictly speaking, to bias our average comparison conclusions, variability of intervals' sizes across individuals is not sufficient: intervals' sizes must differ systematically across the groups in comparison. (See Ravallion *et al.* [2016] for discussion and empirical evidence both about the degree of variability and about how much it affects comparisons.)

This first version amounts to scepticism about intervals' differences of ordinal scales. Yet what could ground this scepticism? Perhaps the idea is that, whenever we are able to form rational beliefs about intervals' differences, it is because we are also able to go all the way and construct a quantitative scale. Thus, the objection says, we never find ourselves only with an ordinal scale and with rational beliefs about intervals' differences.

Arguably, the predominance of RTM in the philosophy of measurement—especially when it is seen as providing an account of how measurement happens in practice—might have made this idea more plausible than it is. In that framework, either one is able to establish a quantitative scale (say, interval) or one is not able to do so, but has only established an ordinal scale. There are no types of scales in between, because there are no axiomatizations in between. However, this doesn't speak to researchers' epistemic situation, that is, to their confidence in their measurements and the inferences they afford (Kyburg [1984]; Tal [2012], pp. 73-78).[14]

From an epistemic perspective, this objection assumes a dichotomy between, on the one hand, knowing nothing about intervals' differences and, on the other, knowing everything (so as to have a quantitative scale). Yet this seems implausible as a description of the situation of researchers at all the stages of measurement development (Larroulet Philippi [2021b]). As statisticians Abelson and Tukey ([1959], p. 226) put it, "the typical state of knowledge short of metric information is not rank-order information; ordinarily, one possesses something more than rank-order information." Other authors have also emphasized that several social science scales may well be more

[14] That RTM isn't a plausible epistemology of measurement is persuasively argued by Sherry ([2011]), Tal ([2012]), among others. Contemporary scholars doubt that RTM is meant to be one (Baccelli [2020]).

informative than ordinal without being quantitative (Labovitz [1967]; Velleman and Wilkinson [1993]; Zumbo and Zimmerman [1993]).

Think of rough measurement procedures, and of unsettled background theoretical knowledge. They may well provide some justification for claims about intervals' differences, one that may allow us to rule out only some (perhaps extreme) intervals' differences—say, that the increase in intelligence between two specific subsequent levels is not ten times larger than the increase in intelligence between other two subsequent levels—while not being enough to allow us to be certain of the intervals' exact sizes. Indeed, this dichotomic assumption is at odds with the gradual accumulation of evidence and development of theory required for justifying the quantitative status of our measurements (Chang [2004]; Sherry [2011]). Well-established quantitative measurement is a hard-won achievement. Thus, there is plenty of space for being on the way to achieve it without already having achieved it.

This dichotomic assumption is also revealed untenable when one considers how measurement claims get their justification from their coherence with background knowledge. To see this, picture the following scenario. Sometimes we may know that a hypothesis such as H holds. For instance, I may know that a group of people under some circumstances are happier on average than another group. Indeed, plausibly, I know that a specific group of people who are struggling to survive in a context of violent civil wars is less happy than another group of people who are living a rewarding life in areas with sustained peace. If we measure the happiness of these two groups, and FOSD is not satisfied, then it follows that the Interval-Difference-K particular to this comparison—whichever it is—holds. By looking at the measurement results I can know which is the K at stake. Since I would know that this particular K holds, I would have rational beliefs about intervals' differences. Yet these beliefs fall short of those involved in a quantitative scale.

All said, the first epistemic objection appears implausible. Perhaps there are many cases in which we cannot have rational beliefs about intervals' differences. But this need not be always the case. So the general prohibition doesn't survive.

4.2 Confining the prohibition

A second objection is less ambitious. It grants that researchers can hold rational beliefs about intervals' differences. But it stresses that ordinal scales are defined as those that inform of rank-order, nothing more (Stevens [1946]). Thus, by definition, beliefs about intervals' differences cannot be grounded on (or represented in) such scales. So the prohibition survives if understood more stringently: as prohibiting making inferences from averages taken with ordinal scales when all information we have comes from the scale. Conclusions about relative averages don't follow from ordinal scale measurements alone (when FOSD doesn't hold), the objection concludes. They follow, when they do, from ordinal scale measurements plus further sources of evidence. Those further sources of evidence account for the beliefs about intervals' differences.

This objection reveals a core aspect of the traditional understanding of measurement scales, as presented by Stevens first, and then by RTM. In this approach, the numerical assignments involved in measurement are meant to only represent a relation (e.g., longer-or-equal-than, at-least-as-happy-as, etc.) that is verified by a straightforward procedure (comparing rods, answering a closed survey question).[15] They do not aim at representing researchers' best estimates—based on theoretical, modelling, and empirical considerations—about the differences in magnitudes between the different objects. This is why from this perspective, the permissible transformations for ordinal

[15] More precisely, under RTM's axiomatization, the numerical assignments are said to represent—by being iso/homomorphic to—an empirical relational system, which consists on the set of objects measured (say, rigid rods, research subjects) and the relation that holds among them (Suppes and Zinnes [1963]).

scales are all the order-preserving ones. Not because every such transformation is (necessarily) compatible with researchers' best judgments about the relative sizes of the intervals all things considered. But because the scale is not meant to capture such judgments; it is only meant to capture the ordering relation directly verified by the procedure.

This understanding of measurement scales (and the coarse-grained classification it gives rise to) faces a problem. This strict definition of ordinal scales—scales that inform of rank-order only—does not characterize well many of the actual scales developed by researchers and that are typically considered ordinal. The term 'ordinal' is used more capaciously than in this strict sense. And this is the case not only when it comes to practitioners. As I argue in subsection 4.3, the concrete examples of ordinal scales given by RTM scholars themselves don't fit the strict definition: these scales aim at being, and arguably are, more informative than rank-orders, despite not being quantitative. The traditional approach to measurement overlooks this because it doesn't even conceive of scales that inform of more than order yet are not quantitative.

So, what shall we say about this second objection? We should grant it: when we are working with strictly defined 'ordinal scales' (that is, those that inform of rank-order only), unless we have some further knowledge about intervals' relative sizes, we cannot make the average comparisons at stake. Indeed, if we interpret 'further knowledge' in the previous sentence undemandingly (say, as being able to rule out at least some intervals' differences), this second objection is just a corollary of my argument in section 3—we are justified in making average comparisons (when FOSD doesn't hold) only if we are justified in believing that the particular Interval-Difference-K at stake holds. (Recall, K is that relation between intervals' differences that is necessary and sufficient in a given case for the hypothesis about relative averages to hold.) Whenever we are not justified in believing that K holds, it just follows from my argument that we cannot infer average comparisons.

Where does this leave us? Recall that the target of my argument (as outlined in the introduction) is the general prohibition against average comparisons using non-quantitative scales. The one that says: ‘unless you have a quantitative scale, you cannot make average inferences.’ This prohibition I have successfully challenged in section 3, and the second objection does not refute my argument there (rather, it is a corollary of my argument). But the point is not merely academic. Indeed, we might call the prohibition targeted here the ‘actual prohibition’. After all, it is the actual prohibition the one that matters for research practice, the one that actually gets deployed in methodological disputes, since there the adjective ‘ordinal’ is used capaciously (for non-quantitative scales in general). Methodologists find doubting the quantitative status of a measurement enough for invoking the prohibition, that is, they don’t feel the need to argue that the scale is strictly ordinal (for example, Bond and Lang [2019], p. 1631).

Hence, as a matter of logic, the actual prohibition is not vindicated by the second objection—since there can be scales that inform of more than order yet are not strictly speaking quantitative. Indeed, that some scales typically labelled ‘ordinal’ are more informative than order I argue below. I focus on the paradigm example, Mohs’ scale. Yet looking briefly at other examples given by RTM scholars of ordinal scales, I’ll suggest that the case of Mohs’ scale is not an exception.

4.3 Mohs’ scale: beyond order

The German mineralogist Friedrich Mohs (1773-1839) worried about the state of his discipline. He had quantitative aspirations for it, and his work on measuring hardness was part of his project to set mineralogy on solid grounds.[16] Mohs ([1825], p. 300) understood hardness-of-minerals as ‘the resistance of solid minerals to the displacement

[16] He also tried to develop a precise mathematical description of how different crystal forms of minerals relate to some basic crystal forms. I thank Steven Irish for sharing his knowledge about Mohs’ program.

of their particles'. In stark contrast with the prevalent view that his 'is the cleanest example of an ordinal scale' (Davis [2018], p. 49), Mohs himself believed his scale to be more informative than rank-orders. Indeed, he believed he had a good sense of how different the intervals were: except for the last (9-10), the intervals are 'proportionate' enough so as to render the scale suitable for the needs of mineralogy ([1825], p. 301). Please pause to note the contrast. We saw above Sherry's remark about Mohs' 'purpose': it was merely to rank-order minerals. Not at all. Mohs ([1825], p. 301) aimed to develop a scale 'capable of ascertaining and indicating [the] differences [in degrees of hardness among minerals], at least with a degree of accuracy and certainty sufficient for the wants of the Natural History of the Mineral Kingdom'. That is why he placed much emphasis on the issue of whether the intervals of his scale were similar enough.

Several studies (summarized by Cambridge physicist David Tabor [1954], [1970]) arguably proved Mohs correct—except for the last interval that is much bigger than the rest, the intervals are not wildly different. These studies employed contemporary techniques for measuring hardness, which use indentation procedures, quantifying the indentation's depth (given some force). The result? If we remove the (correctly diagnosed as) much bigger 9-10 interval, the numbers in Mohs' scale roughly obey the relation log(Hardness)=1.6 Mohs. So, although clearly not equal (each interval is roughly 60% larger than the previous), the intervals of Mohs' scale are not wildly heterogeneous.[17] Think that, in the fictitious case discussed in subsection 3.1, Mohs

[17] Are scratching and indenting procedures measuring the same attribute, hardness-of-minerals? Note Mohs' definition of hardness—it is wide enough to be captured by both procedures. Moreover, as Tabor explains, under suitable conditions both scratching and indenting measure the 'plastic' properties of solids such as minerals and metals. This explains the uniform relation found between scratching and indenting in both metals and minerals ([1954], [1970]).

surely would have deduced hypothesis *H*. And he would have got it right, according to contemporary measurements of hardness (Whitney *et al.* [2007]).

Thus, Mohs believed his scale provided more information than merely order, and he was right. But how was this possible, if all he had is a scratching procedure that, it seems, only produces orderings? This question betrays a misunderstanding of Mohs' scale, one that is arguably made more likely by the widespread endorsement of the traditional view of measurement scales discussed above. Mohs' scale is usually described as the numerical assignment that is consistent with the ordering provided by the scratching procedure (recall Sherry's long quotation). But Mohs' scale does not merely aim at representing the results of a scratching procedure.

For starters, it also involves singling out ten specific minerals (or 'standards') which are assigned specific numbers. Thus, besides the (usually only mentioned) scratching, there is the task of choosing which mineral exemplifies a 1, which one a 2, etc. There is space to consider evidence beyond order here: one needs to choose which mineral, among the potentially many that scratches a 1 and are scratched by a 3, lies roughly in between the 1 and the 3 along the hardness dimension. Mohs ([1825], p. 301) made it clear that we achieve a good (enough) scale by 'choosing a certain number of suitable minerals […] taking care always that the intervals between every two members of the scale be not so disproportionate, as either to render its employment more difficult, or to hinder it altogether'. In other words, the task is to choose minerals that are similarly spaced out along the hardness dimension, so that the scale is useful (that is, more informative than order). Tabor took this judicious selection of minerals as the crucial ingredient in explaining why Mohs did better than merely ordering (as reported above). Tabor ([1954], p. 256) said: 'This regularity in behaviour suggests that Mohs did not simply choose "ten common minerals arranged in order of increasing hardness" ([reference omitted]): it would seem that he experimented with a much larger variety until he had satisfied himself that he had obtained "equality of the intervals"'.

That Mohs experimented much, and with many different minerals, is evident from his description of the scale's construction ([1825], pp. 300-07). For example, he mentions not one but several varieties of the minerals that appear in the scale, making clear which varieties are(n't) of equal hardness;[18] and he mentions several other minerals with the same degree of hardness than those of the scale but that were not chosen because they are less available. But why would this latitude to choose the minerals that go in the scale, coupled with vast experimentation, matter for justifying the claim that Mohs' scale informs more than order? If all what Mohs had was the scratching test, how could he make use of his vast experimentation to infer something beyond order? Tabor ([1954], p. 257) suggests Mohs used a 'tactile criterion' for estimating the differences in hardness, but he doesn't elaborate on the proposal. Perhaps he meant Mohs noted how deep the scratches went, or how easy to scratch was, for different pairs of minerals. All this is possible, and points towards reasonable ways of forming beliefs about intervals' differences. What is clear from Mohs' ([1825], p. 304) account, however, is that he saw the need to rely on more than the 'mere scratching' procedure: 'Numerous experiments of determining the degree of hardness, by the mere scratching of one substance with the other, have completely established, that this process alone is not sufficient'. Besides the scratching procedure, he used a file to gauge how much the minerals resisted and how much noise they made. He experimented until he was satisfied that all the specimens of a given mineral reacted similarly to the file. His extensive experimentation with the file made him confident in the selection of minerals for establishing approximately equal intervals.

[18] For example, after stating which mineral is assigned a 3, Mohs ([1825], p. 302) makes clear that two other varieties of the same family 'cannot be employed in its place, the hardness of these being considerably higher'. Talk of 'considerably' makes clear he thought able to judge differences in degrees beyond order.

Thus, close attention to Mohs' case reveals that his scale doesn't reduce to the 'harder-or-equal-than' relation verified by the scratching procedure. Unfortunately, that is the view most (if not all) methodologists and philosophers have of it, one that fits nicely the traditional understanding of measurement scales, where numerical assignments represent specific relations verified by straightforward procedures. But Mohs' scale combines information from the scratching procedure with other (perhaps less definite) procedures, such as using the file (or tactile criteria) for gauging resistance. And the selection of minerals, which fixes the intervals of the scale, is based on extensive experimentation with these procedures, and draws on background knowledge of the minerals. This is what allows Mohs to justifiably be confident that his scale informs of more than order, without thereby being quantitative. All said, the extant literature misinterprets Mohs' work, and underestimates the epistemic status of his scale.

Let us take a step back. On the face of it, there is nothing special about Mohs' efforts to make his scale more informative than order. Extensive experimentation and drawing upon background knowledge—versus drawing only on a single definite procedure that merely orders—may well characterize the process that goes in the construction of several scales that are typically deemed 'ordinal'. Would it be surprising that much effort is put into achieving scales more informative than order? Psychologists developing self-reporting scales of, say, happiness, might aim at picking answer categories that are roughly equally spaced along (their understanding of) the happiness dimension. They may not all succeed, but it requires much argumentation (yet to be seen) to claim that all necessarily fail. In general, the same procedures that allow researchers to rank elements in terms of an attribute may also allow them to have some sense of the differences between levels, even if they aren't definite enough to afford a strictly quantitative scale. One can think here of academic assessments (or intelligence tests), where the understanding of the subject matter that enables test experts to rank questions in terms of their difficulty may also enable them to have some coarse-grained sense of intervals' differences. In other cases, complementary procedures might come up to play that role (such as using a file, in the case of Mohs, or calibrating happiness self-reporting scales with independent evidence of happiness as was illustrated above).

I briefly comment on two other examples given by RTM scholars. We saw Roberts ([1985]) considering air quality scales as ordinal. He gave no argument for this, nor details of the specific scale. So we cannot fully assess the case. But air quality indices are constructed using the amount of particulate material in the air, averaging (or taking the maximum value of) different quantitative indicators. The total scores are then classified in terms of their danger to our health in few categories. Given that those categories are based on quantitative indicators (about the amount of different particulate material), the scale constructors may have some sense of the distances between the categories. Or, at the very least, they are not totally ignorant about intervals' differences.

Another example mentioned is the Beaufort Wind Force Scale (Suppes and Zinnes [1963]; Baccelli [2020]). There is no 'one' Beaufort scale, but rather an evolution of scales. Nevertheless, here again the claim that the scale is strictly speaking ordinal is untenable. The differences in average wind speed of each pair of subsequent categories, as were measured long ago, are not exactly equal. But they are far from being wildly heterogeneous.[19] Again, though the scale is not quantitative, it doesn't leave its users totally at loss about intervals' differences.

Summing up, classical examples given of ordinal scales by RTM scholars don't fit the strict definition of an ordinal scale. These scales are more informative than order, without being quantitative. This is what makes rejecting the actual prohibition an important step forward in the methodological disputes around measurement scales.

[19] For the historical evolution of the scale and the actual wind velocities see Met Office (n.d.).

Before concluding, I consider a pragmatic objection: it is better to have a clear (though admittedly too strict) rule than to have no rule, leaving the judgment of whether a non-quantitative scale may be used for an average comparison in any given case to researchers themselves.[20] As an analogy to motivate this objection, think that you don't need to fully endorse frequentist statistics to defend a clear and fixed convention for reporting results (say, a p-value of 0.05). You may think that such rule is less than epistemically optimal in principle, but better in practice. Scientists, after all, are not the ideal agents modelled by Bayesians.[21]

Although an important point, and one which calls for detailed analysis using tools from social epistemology, I'll note the following (without pretending to settle the issue). Sticking blindly to the prohibition is not analogous to taking a fixed middle position that avoids arbitrariness at the cost of being too conservative sometimes and too risky others, as (arguably) in the case of the fixed-p-value rule. Rather, it amounts to being as conservative as possible in all circumstances, no matter how close to being quantitative some scales typically deemed ordinal may be. Second, recall that adhering to the prohibition dramatically limits the research that the social and medical sciences can do. Taking these two points together, I think it is doubtful that this rule would be 'better in practice'.

5 Conclusion

I have challenged the general prohibition against making inferences from averages taken with non-quantitative scales. Average comparisons don't require quantitative scales: the beliefs needed to justify these inferences are less demanding. These beliefs, which concern the relative differences of specific (versus all) intervals, may well be held by researchers working with scales usually considered ordinal. This is arguably the case

[20] I thank Anna Alexandrova for raising this objection.

[21] The analogy is inspired by (Steele [2012]).

for the purported paradigm ordinal scale, Mohs' scale of hardness of minerals, but this may also hold for other cases.

It doesn't follow that all scales considered ordinal are suitable for inferences using averages. As bad as a general prohibition that assumes that all non-quantitative scales inform only about order is to assume that the warrant for beliefs about intervals' differences is the same across all these scales. There may well be cases for which talk of intervals' differences indeed does not make much sense. Cases, that is, where scale developers don't even aim at anything more than ordering. (House numbering or lists ordered alphabetically come to mind; but compare the ordering of students induced by their names with the one induced by their numerical grades. Most teachers that I know of are confident in that some grades inform of more than order.) And there may well be cases for which our total evidence is extremely thin, so that little (if any) confidence should be placed in average comparisons. However, these need not be true of all cases. In short, scales typically considered ordinal need not be all on a par with regard to their informativeness beyond order. Thus, the justification for using averages needs a case-by-case assessment. Neither a general prohibition nor anything goes are defensible positions.

Acknowledgments

Thanks to Anna Alexandrova, Hasok Chang, Mahmoud Jalloh, Miguel Ohnesorge, Raul Saucedo, Julia Staffel, Jacob Stegenga, two referees, and participants at the British Society for Philosophy of Science Conference (2021), the Biennial Meeting of the European Philosophy of Science Association (2021), and the Annual Conference of the Chilean Philosophy Association (2021) for helpful comments and suggestions. This paper is based on research funded by the Gates Cambridge Trust; Chilean National Research and Development Agency (Becas Chile).

Cristian Larroulet Philippi
Department of History and Philosophy of Science
University of Cambridge
Cambridge, UK
cristianlarroulet@gmail.com

https://orcid.org/0000-0001-5793-4670

References

Abelson, R. P. and Tukey, J. W. [1959]: ‘Efficient Conversion of Non-Metric Information into Metric Information’, *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 226-30.

Acemoglu, D., De Feo, G. and De Luca, G. D. [2020]: ‘Weak States: Causes and Consequences of the Sicilian Mafia’, *The Review of Economic Studies*, **87**, pp. 537–81.

Baccelli, J. [2020]: ‘Beyond the Metrological Viewpoint’, *Studies in History and Philosophy of Science*, **80**, pp. 56–61.

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

Cetre, S., Clark, A. and Senik, C. [2016]: ‘Happy People have Children: Choice and Self-Selection into Parenthood’, *European Journal of Population*, **32**, pp. 445-73.

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

Davis, R. [2018]: ‘Why the Mohs Scale Remains Relevant for Metrology’, *IEEE Instrumentation & Measurement Magazine*, **21**, pp. 49-51.

Easterlin, R. [1974]: ‘Does Economic Growth Improve the Human Lot?’, in P. David and M. Reder (*eds*), *Nations and Households in Economic Growth:*

Essays in Honor of Moses Abramovitz, New York: Academic Press, pp. 89-125.

Frijters, P. and Krekel, C. [2021]: *A Handbook for Wellbeing Policy-Making,* Oxford: Oxford University Press.

Hadar, J. and Russell, W. [1969]: 'Rules for Ordering Uncertain Prospects', *The American Economic Review*, **59**, pp. 25-34.

Haushofer, J. and Shapiro, J. [2016]: 'The Short-Term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya', *Quarterly Journal of Economics,* **131**, pp. 1973-2042.

Howson, C. and Franklin, A. [1994]: 'Bayesian Conditionalization and Probability Kinematics', *The British Journal for the Philosophy of Science,* **45**, pp. 451-66.

Krantz, D., Luce, D., Tversky, A. and Suppes, P. [1971]: *Foundations of Measurement*, Mineola: Dover.

Krosnick, J. A. and Presser, S. [2010]: 'Question and Questionnaire Design', In P. V. Marsden and J. D. Wright (*eds*), *Handbook of Survey Research,* Bingley: Emerald Group Pub., pp. 263-313.

Kyburg, H. [1984]: *Theory and Measurement*, Cambridge: Cambridge University Press.

Kyburg, H. [1992]: 'Measuring Errors of Measurement', in W. Savage and P. Ehrlich (*eds*), *Philosophical and Foundational Issues in Measurement Theory,* New Jersey: Lawrence Erlbaum Associates, pp. 75-91.

Larroulet Philippi, C. [2021a]: 'Valid for What? On the Very Idea of Unconditional Validity', *Philosophy of the Social Sciences,* **51**, pp. 151-75.

Larroulet Philippi, C. [2021b]: 'On Measurement Scales: Neither Ordinal nor Interval?', *Philosophy of Science,* **88**, pp. 929-39.

Labovitz, S. [1967]: ‘Some Observations on Measurement and Statistics’, *Social Forces*, **46**, 151-60

Lord, F. [1953]: ‘On the Statistical Treatment of Football Numbers’, *The American Psychologist*, **8**, pp. 750–1.

Ludwig, J., Duncan, G., Gennetian, Katz, L., Kessler, R., Kling, J. and Sanbonmatsu, L. [2012]: ‘Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults’, *Science,* **337**, pp. 1505-10.

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

Met Office. [n.d.]: *National Meteorological Library and Archive Fact Sheet 6 — The Beaufort Scale*. Available at <https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/research/library-and-archive/library/publications/factsheets/factsheet_6-the-beaufort-scale.pdf>.

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. [2012]: ‘“The Constantly Recurring Argument”: Inferring Quantity From Order’, *Theory and Psychology*, **22**, pp. 255-71.

Mohs, F. [1825]: *Treatise on Mineralogy, Or, the Natural History of the Mineral Kingdom. Tr. from the German, with Considerable Additions, by William Haidinger*, Vol. 3, Edinburgh: Archibald Constable and Co.

Ravallion, M., Himelein, K. and Beegle, K. [2016]: ‘Can Subjective Questions on Economic Welfare Be Trusted?’, *Economic Development and Cultural Change*, **64**, pp. 697-726.

Roberts, F. [1985]: *Measurement Theory,* Cambridge: Cambridge University Press.

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

Steele, K. [2012]: ‘The Scientist qua Policy Advisor Makes Value Judgments’, *Philosophy of Science,* **79**, pp. 893-904.

Stevens, S. S. [1946]. ‘On the Theory of Scales of Measurement’, *Science*, **103**, pp. 667-80.

Suppes, P. and Zinnes, J. [1963]: ‘Basic Measurement Theory’, in R. D. Luce, R., R. Bush and E. H. Galanter (*eds*), *Handbook of Mathematical Psychology*, Vol. 1, New York: Wiley, pp. 1-76.

Tabor, D. [1954]: ‘Mohs's Hardness Scale—A Physical Interpretation’, *Proceedings of the Physical Society. Section B,* **67**, pp. 249-57.

Tabor, D. [1970]: ‘The Hardness of Solids’, *Review of Physics in Technology,* **1**, pp. 145-79.

Tal, E. [2012]: *The Epistemology of Measurement: A Model-Based Account*, PhD Thesis, University of Toronto.

Tal, E. [2017]: ‘Measurement in Science’, in E. N. Zalta (*ed*), *Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/archives/fall2017/entries/measurement-science/>.

Tennant, A., McKenna, S. P. and Hagell, P. [2004]: ‘Application of Rasch Analysis in the Development and Application of Quality of Life Instruments’, *Value in Health,* **7**, pp. S22-S26.

Whitney, D. L., Fayon, A. K., Broz, M. E. and Cook, R. F. [2007]: ‘Exploring the Relationship of Scratch Resistance, Hardness, and Other Physical Properties of Minerals using Mohs Scale Minerals', *Journal of Geoscience Education*, **55**, pp. 56-61.

Velleman, P. and Wilkinson, L. [1993]: 'Nominal, Ordinal, Interval, and Ratio Typologies are Misleading', *The American Statistician*, **47**, pp. 65-72.

Zumbo, B. and Zimmerman, D. [1993]: 'Is the Selection of Statistical Methods Governed by Level of Measurement?', *Canadian Psychology/Psychologie Canadienne*, **34**, pp. 390-400.