Philosophy of Science (2025), **92**, 1205–1214
doi:10.1017/psa.2025.10173

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

CONTRIBUTED PAPER

Measured Inference: Scales, Statistics, and Scientific Inference

Conor Mayo-Wilson

Department of Philosophy, University of Washington, Seattle, WA, USA
Email: conormw@gmail.com

(Received 12 January 2025; revised 19 August 2025; accepted 19 August 2025; first published online 04 September 2025)

Abstract

Despite the recent "epistemic turn" in the philosophy of measurement, philosophers have ignored a nearly 80-year controversy about the relationship between statistical inference and measurement theory. Some scholars maintain that measurement theory places no constraints on statistics, whereas others argue that the measurement scale (e.g., ordinal or interval) of one's data determines which statistical methods are "permissible." I defend an intermediate position: Even if existing measurement theory were irrelevant to statistical inference, it would be critical for *scientific* inference, which requires connecting statistical hypotheses to broader research hypotheses.

1. Introduction

Despite the recent "epistemic turn" in the philosophy of measurement, philosophers have ignored a nearly 80-year controversy about the relationship between statistical inference and measurement theory.[1] Statistical libertarians, as I will call them, maintain that measurement theory places essentially no constraints on statistics.[2] In contrast, measurement bureaucrats (again, my term) endorse Stevens's doctrine of permissible statistics, according to which parametric methods (e.g., t-tests) should be applied only to interval or ratio-scaled data, whereas ordinal data require the use of nonparametric tests (e.g., Mann–Whitney U).[3]

[1] See Tal (2015) for a discussion of the "epistemic turn." To my knowledge, the only philosophical works that engage with this controversy are those by Larroulet Philippi (2021, 2022). As is common (e.g., Tal 2015), I use the term *measurement theory* to refer to the mathematical work that culminates in the three-volume *Foundations of Measurement* texts (Krantz et al. 1971). I avoid using the term *representational measurement theory* (RTM) to describe those mathematical results because RTM is often used to denote several further epistemological theses Tal (2021).

[2] See Lord (1953, 1954), Gaito (1980), Atkinson (1988), and Velleman and Wilkinson (1993).

[3] Bureaucrats include Blalock (1960), Wilson (1971), Senders (1958), Siegel and Castellan (1988), and Thomas (2006). An intermediate position is defended by Marcus-Roberts and Roberts (1987), who argue

© The Author(s), 2025. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

To see what is at stake, imagine a chief executive officer (CEO) is worried about sexual harassment in her company. She issues a two-question survey to 50 randomly selected employees. The first question asks employees to identify their gender, and the second asks, "On a 1–7 scale, with 1 representing 'completely dissatisfied' and 7 representing 'completely satisfied,' how satisfied are you with the company's sexual harassment policies?" When the survey has been completed by all 50 selected employees, the CEO divides responses according to gender and calculates the averages/means of men's and women's responses (3.2 and 2.1, respectively). She then performs a t-test to assess whether those averages differ. She finds a statistically significant difference. May the CEO conclude that men and women in the company are satisfied to different degrees with the company's sexual harassment policies?

Libertarians maintain "yes"; bureaucrats say "no." According to libertarians, if the *averages* of the men's and women's responses differ, then so must the two distributions of responses. End of story.

For bureaucrats, however, the CEO's data are merely *ordinal*, and averages of ordinal data should not be invoked in statistical reasoning. To motivate that prohibition, suppose the survey had omitted a numerical scale and instead asked respondents to choose from seven categories describing their degree of satisfaction in English words. Just as a researcher might be hesitant to "average" nonnumerical responses like "somewhat dissatisfied" and "very satisfied," one should be reluctant to calculate the means of the responses from the original survey.

I defend an intermediate position: Even if existing measurement theory were irrelevant to statistical inference, it is critical for *scientific* inference, which requires connecting statistical hypotheses to broader research hypotheses.[4] In the CEO's case, the statistical hypotheses concern the relationship between two (probability) distributions over the numbers 1–7, which represent employee responses on a fixed numerical scale. In contrast, the research hypothesis of interest likely concerns whether *attitudes* about sexual harassment differ or whether men's and women's *behavior* differ in ways that matter to the CEO (e.g., whether productive women are more likely to leave the company within the year). Libertarians are correct that the CEO's statistical inferences require no measurement theory, but bureaucrats are correct that further assumptions are necessary to draw research conclusions from the CEO's statistical analysis.

To show how the distinction between statistical and research hypotheses arises in scientific practice, I summarize the controversy on "interpretable effects" in psychology in Section 2.[5] I argue that the controversy amounts to the following: Statistical conclusions reached in memory experiments do not *mathematically entail*[6] research hypotheses about some purported latent attributes, specifically, an attribute

that only "meaningful" statistical hypotheses are of scientific interest (see sec 3.2) but that there are no restrictions on what statistics it is appropriate to calculate.

[4] The distinction between statistical and research hypotheses is standard in medical science. See Lawler and Zimmermann (2021) for a discussion of cases in which the two types of hypotheses are misaligned.

[5] The controversy originated with Loftus (1978). See Wagenmakers et al. (2012) for the history.

[6] Henceforth, I say a set of premises *mathematically entails* a conclusion if the premises of the argument and the axioms of set theory together *logically* entail the conclusion. I say an argument is *mathematically valid* if its premises mathematically entail its conclusion.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

that might be "memory strength." Moreover, many psychologists believe that the data from some memory experiments are of questionable scientific interest unless the statistical hypotheses that the data support mathematically entail the relevant research hypotheses.

I remain agnostic regarding what is of "scientific interest" in psychology, but I investigate the consequences of the skeptical position about memory experiments. In Section 3, I argue that the theory of "meaningfulness" developed in measurement theory can help one identify general conditions under which inferences from statistical hypotheses to research ones are mathematically valid.[7]

Before beginning, it will be helpful to characterize the distinction between statistical and research hypotheses more precisely. Statistical hypotheses are (sets of) *probability distributions* that specify how likely various data are. Such hypotheses always concern a *particular* experimental setup (which might be repeatable). Statistical methods (e.g., hypothesis tests and estimators) allow one only to evaluate how well different statistical hypotheses are supported by data.

In contrast, research hypotheses have implications beyond a given experimental context, and they may not specify any precise probabilities whatsoever. A research hypothesis might, for example, concern (i) latent or unmeasured attributes or (ii) the outcomes of different measurement procedures in other experimental contexts. In the CEO's case, the latent attributes are attitudes or behavioral dispositions, which are not measured in the survey.

2. Conflating latent attributes with measured ones

Nearly 50 years ago, Loftus (1978) famously argued that many memory experiments in psychology suffer from a serious methodological problem: A latent attribute—call it *memory strength*—is conflated with what is directly measured in experiments, for example, the probability of correctly recalling a stimulus.

Imagine experimental subjects are divided into two groups; call them A and B. Participants in both groups are presented with a sequence of five "random" letters, which they will be asked to recall at two different later times (e.g., after 5 and 20 seconds, respectively). But prior to the recall phase of the experiment, different groups are subject to different conditions. Groups A and B might receive different instructions, for example.

Suppose the results of the experiment are as shown in figure 1. Group A's average recall rates are represented by the two circular end points of the bottom line, and group B's average recall rates are represented by the two square end points of the top line.

The lines in the diagram are merely heuristic. Both groups are tested for recall at only two discrete times, so the lines do *not* indicate that, in the experiment, the probability of correct recall decreases linearly over time. However, the lines help one see an important fact: The slope of the group A line is steeper than that of the group B line. One might hypothesize, therefore, that participants in condition A *forget* at a

[7] I use the theory of *semantic* meaningfulness noted by Adams et al. (1965). This theory was inspired by the theory of "empirical" or "scientific" meaningfulness that was developed by Suppes and Zinnes (1962) and later defended by Roberts (2009) and Narens (2012), among others. I do not endorse the latter theories.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

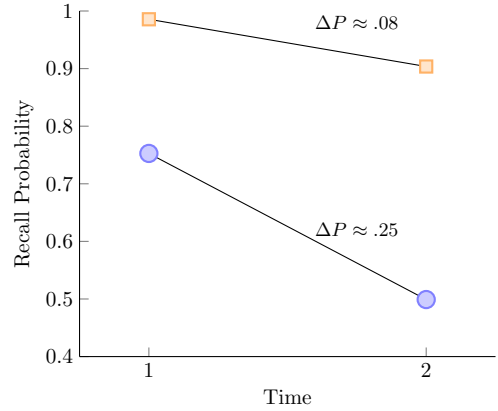


Figure 1. Recall Probability vs. Time.

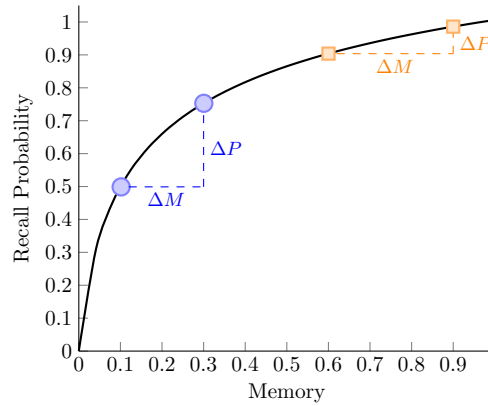


Figure 2. Recall Probability vs. Memory.

faster rate than do participants in condition B. That hypothesis, Loftus argues, is underdetermined by the experiment.

Why? Suppose memory strength—call it q—is quantifiable, and suppose observed recall rate at time t is a function r_t of q. Loftus shows that even if recall rate increases with q, it is possible for q to decrease *at the same rate* or even *faster* in condition B than in condition A, unless one makes further assumptions about the mathematical form of the function r_t. Which assumptions? Loftus proves that if the recall rate is a *linear* function of q, then the desired inference about memory strength is valid.

Figures 1 through 3 illustrate Loftus's critique. Suppose the function r_t is the one shown in figure 2. Then *memory* decreases faster in condition B than in condition A (as shown in figure 3), even though *recall rates* in condition B decrease more slowly than in condition A (as shown in figure 1).

Loftus never distinguishes between statistical and research hypotheses, but his critique beautifully illustrates the distinction. Even if statistical methods establish that the distribution of *observed recall rates* in condition A differs from that in B,

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

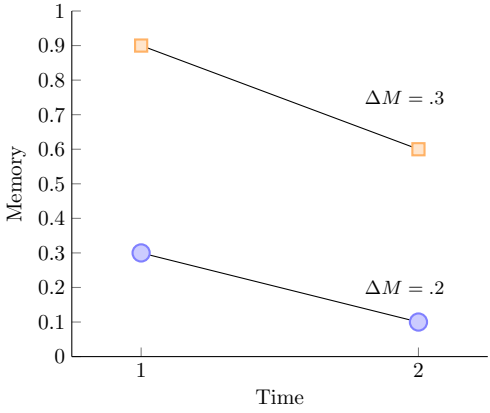


Figure 3. Memory vs. Time.

Loftus's critique challenges the inference from those statistical conclusions to research hypotheses about memory.

One might object to Loftus's critique by arguing that in the memory experiments in question, psychologists are not trying to draw inferences about a latent attribute: The research hypotheses and statistical hypotheses alike concern the measured probability of recall. Memory is "operationalized" in recall rates. That critique would be legitimate if such recall rates were known to be of independent scientific interest. For example, perhaps those recall rates can predict performance in many other important "memory" tasks. But crucially, the *differences* in recall rates must be predictive because otherwise, the conclusion that subjects' recall rates *decrease* faster in one condition than in another is irrelevant.

3. Inference, meaningfulness, and scales

Loftus shows that in important scientific settings, there may be an inferential gap between research hypotheses about latent attributes and statistical hypotheses about measured outcomes. I now argue that in many of those settings, the theory of meaningfulness developed by measurement theorists specifies the assumptions necessary to bridge the inferential gap. To do so, I first argue that the theory of meaningfulness provides a plausible answer to the question, "Under what conditions does an attribute (e.g., memory strength) have the type of quantitative structure for which differences (e.g., between memory strength at two different times) are meaningful?" The answer, I claim, is that the attribute admits an *interval scale*.[8] Similarly, claims about ratios of an attribute are meaningful if and only if the attribute admits a ratio scale. I discuss scale types in section 3.1.

My main contribution is to show that scale classifications also play an important *epistemic* role because they can be used to identify *mathematically valid* inferences from statistical hypotheses to research ones. Because inferences from statistical

[8] Because I am not a psychologist, I will not assess whether there is evidence that "memory strength" admits an interval scale.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

hypotheses to research ones are often not easily formalized, it is important to identify which are mathematically valid.

3.1. Scales and scale types

Length can be quantified in inches and centimeters. Mass is quantified in kilograms and tonnes. In general, anything that is quantifiable can be quantified in many ways. Roughly, a *scale* is a way of quantifying a property. Scales are rarely unique.

But scales are often related. For example, an inch is 2.54 centimeters; a yard is 3 feet, and generally, one can convert any unit of length into another by multiplying by a constant. When all scales for an attribute are multiples of one another, the attribute is called *ratio scaled*.

Not all scales are ratio scales. Consider calendar date. In all calendar systems, there is an arbitrarily chosen "zeroth" year, and calendar date is determined by counting from that zero. Different zeroes can be chosen; for example, in Islamic calendars, Muhammad's pilgrimage fixes the zeroth year. And instead of counting years, one could count days, weeks, or units of time determined by lunar rather than solar events. Thus, in converting calendar date in one system to another, one must first multiply (e.g., to convert years to days) and then add another number (e.g., to correct for choices of "zeroth" year). When all scales for a property are related in this way, one says the property is *interval scaled*.

The reader might ask, "What determines which scales are 'permissible' ways of quantifying an attribute?" For the purposes of this article, my answer is, "Consult *Foundations of Measurement*" (Krantz and Tversky, 2006, Krantz et al., 1971, 2006).There, the reader will find a body of mathematical theorems showing that if certain relations hold among objects (or events) with a given attribute, then the set of permissible scales must always be of one of the few types identified by Stevens (1946). Importantly, as Michell (1997) observes, the theorems in *Foundations of Measurement* specify the quantitative structure of an attribute even if the attribute cannot be measured in any realistic sense. This is important because Krantz et al. (1971) are often said to endorse "positivist" assumptions, for example, that "empirical relations be directly observable, or 'identifiable'" (Mari et al., 2023, p. 94). Those philosophical assumptions, however, play no role in the mathematical results about scale types.

What is now important for us is to understand how scale classifications can clarify questions about meaningfulness.

3.2. Meaningfulness

Contrast two claims: "Ada is more than twice as tall as Boris" and "Ada's height in inches is more than twice that of Boris." Notice that the first sentence is true if and only if the second is true. That may be surprising because the first is a *scale-free* assertion—it contains no mention of *units* of length—whereas the latter is *scale specific*. But according to an influential definition of "meaningfulness," one should not be surprised at all: The first sentence has a truth value if and only if its truth value matches that of the second.[9]

[9] See references in footnote 3.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

To understand the proposed theory of meaningfulness, consider the scale-free assertion "Ada's is three taller than Boris." That claim is nonsense. If Ada is 3 inches taller than Boris, then she is not 3 feet taller than Boris. Units matter. These examples motivate the following proposal: A scale-free sentence about an attribute is *meaningful* (i.e., it has a truth value) if and only if all the scale-specific instances of the statement have the same truth value. In other words, a scale-free sentence is meaningful if the units do not matter.

Scale-free hypotheses are ubiquitous in science. Consider Galileo's law of free fall, which asserts that the distance traveled by an object in free fall is proportional to the square of the time of the descent. Galileo's law does not require that distance be measured in a specific unit, such as meters, nor that time be measured in a unit such as seconds. Similarly, Boyle's law about pressure and volume is scale-free: Neither units of pressure nor units of volume are mentioned. Scale-free hypotheses also occur in the social sciences. For example, economists do not mention a specific currency when they claim that profits are maximized when marginal revenue equals marginal costs. These examples show that it is important to understand when scale-free hypotheses are meaningful.

To see how the theory of meaning works, consider the scale-free hypothesis "Memory strength decreases more rapidly in condition A than in condition B between times t_1 and t_0." Loftus argued that the hypothesis could not be inferred from the observed recall effects.

However, that scale-free hypothesis is meaningful, according to the previously described theory of meaning, if for any two scales for memory M_1 and M_2, the following biconditional holds:

$$M_1(t_1, A) - M_1(t_0, A) > M_1(t_1, B) - M_2(t_0, B) \text{ if and only if}$$
$$M_2(t_1, A) - M_2(t_0, A) > M_2(t_1, B) - M_2(t_0, B), \quad (1)$$

where $M_j(t, x)$ represents the memory strength along scale $j \in \{1, 2\}$ at recall time $t \in \{1, 2\}$ in condition $x \in \{A, B\}$. Some quick algebra shows that equation 1 holds if there is a positive number $c > 0$ and some number d (possibly negative) such that $M_2(t, x) = c \cdot M_1(t, x) + d$ for all times t and all conditions x. That is, the assertion is meaningful if memory is an interval-scaled attribute.

This example suggests that there is some relationship between (1) meaningfulness and (2) the validity of inferences that have scale-free conclusions. Understanding that relationship is important because whereas *statistical* hypotheses are almost always scale specific (because they describe the data of a particular experiment, which must be measured in specific units), scientists' *research* hypotheses are often scale-free.

3.3. Mathematical validity and research hypotheses

The theory of meaningfulness allows us to immediately identify a set of mathematically valid inferences that have scale-free conclusions. Let M be a scale; let φ_M be some scale-specific proposition about the attribute A, and let φ_A be the corresponding scale-free proposition. For instance, if M is inches and φ_M is the assertion "Ada's height in inches is twice that of Boris," then φ_A is the assertion "Ada's height is twice that of Boris." Here's a theorem (stated imprecisely):

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

Fact: The inference from φ_M to φ_A is valid if (1) φ_A is meaningful, and (2) M is a permissible scale for the attribute A.

The fact follows immediately from definitions. Suppose 1 and 2 hold. Because φ_A is meaningful (by 1), φ_A is true if and only if φ_S is true for any scale S. Because M is a scale for A (by 2), it follows that if φ_M is true, then φ_A must be true (and so the inference is valid).

So what? Recall that statistical hypotheses are about data in a given experimental context *on a fixed scale* (e.g., the CEO's data are on a 1–7 scale for satisfaction; the memory experiment's data are the probability of recall in a specific context). In contrast, research hypotheses are often scale-free precisely because researchers desire replicable results that do not depend on the choice of measurement units. Thus, the inference from a statistical hypothesis (e.g., that men's and women's responses differ on average) to the corresponding scale-free research hypothesis is mathematically valid if (1) the research hypothesis is meaningful, and (2) the measurement scale is a permissible way of quantifying the attribute.

This simple fact is a generalization of Loftus's positive suggestion. It entails that if memory strength admits an interval scale (and so the hypothesis that memory decreases faster in one condition than another is meaningful), then one can validly infer the research hypothesis from the measured results about recall rate if the recall rate is a permissible scale for memory—that is, it is a linear function of memory strength. However, this fact is a generalization of Loftus's claim because it applies to *all* scale types, not just interval ones.

The epistemological importance of this fact, however, should not be overstated. Notice that in Loftus's critique—as in the previously stated fact—(1) the focus is on the *validity* of arguments rather than inductive strengths, and (2) the conclusion of the inference φ_A is the scale-free hypothesis *corresponding* to the premise φ_M. That is a very restricted form of inference.

First, many strong arguments are not mathematically valid. Second, many valid inferences are not of the previously described form and yet have scale-free conclusions. For instance, let ψ and φ be scale-specific and scale-free hypotheses, respectively. Suppose φ is meaningful. Then $\psi \to \varphi$ and ψ together entail φ.

In fact, it is possible to describe such a case of modus ponens when the scale of ψ is not even a permissible scale for the relevant attribute. Suppose two cross-country teams race one another. Let φ be the (scale-free) hypothesis that asserts, "The average time of team 1 is faster than that of team 2." Notice that the hypothesis is meaningful because its truth does not depend on whether times are recorded in seconds, milliseconds, and so forth. However, suppose the finishing times of runners are not recorded, only the ranks, with 1 being assigned to the first-place runner, 2 to the second-place runner, and so on. The assignment of ordinal ranks is not a permissible scale for time. But let ψ be the scale-specific proposition "All runners on team 1 have a lower rank than all runners on team 2." Then $\psi \to \varphi$ is a mathematical truth, so if one has evidence for the scale-specific claim ψ, then one obtains evidence for the scale-free hypothesis φ.[10]

[10] More generally, this argument would work if ψ were replaced with the claim that the ranks of team 2 stochastically dominate those of team 1.

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

Despite these limitations, the theory of meaningfulness provides a first step in (i) understanding the debate between statistical libertarians and measurement bureaucrats and (ii) identifying a partial resolution.

4. Conclusion

As others have convincingly argued, measurement theory helps scientists identify whether an attribute is quantifiable at all.[11] I have further argued that if conditions for quantifiability are met, measurement theory characterizes auxiliary assumptions that are sufficient to facilitate mathematically valid inferences from statistical hypotheses about measured outcomes to research hypotheses about latent attributes. Namely, by the simple fact established in Section 3.3, it suffices to show that the measured outcomes are values along a permissible scale for the attribute.

If the measurement scale is not a permissible (or if there is no latent attribute with the relevant mathematical structure), then often, further data and statistical analyses will be necessary to facilitate inference to research hypotheses. This is the most plausible way of describing the CEO case at the outset of this article. "Satisfaction with sexual harassment policies" is likely not a latent attribute admitting an interval scale, and what researchers are likely interested in is making inferences from the survey to other behaviors. But such inferences would require data that would allow one to explore statistical associations between survey responses and the relevant behaviors.

This article has only begun to address the question of when scale-specific propositions provide *evidence* for scale-free ones. I have discussed only a very narrow set of mathematically valid inferences, and a general theory of *inductive* inference for scale-free hypotheses is still in its infancy (Larroulet Philippi, 2021, 2022).

Acknowledgments. Thanks to Cristian Larroulet Philippi, audience members at the Philosophy of Science Association, and especially to Paul Pedersen and David Kellen for earlier conversations about measurement and statistics.

References

Adams, Ernest W., Robert F. Fagot, and Richard E. Robinson. 1965. "A Theory of Appropriate Statistics." *Psychometrika* 30 (2):99–127. https://doi.org/10.1007/BF02289443.

Atkinson, Leslie. 1988. "The Measurement-Statistics Controversy: Factor Analysis and Subinterval Data." *Bulletin of the Psychonomic Society* 26 (4): 361–64. https://doi.org/10.3758/BF03337683.

Blalock, Hubert. 1960. *Social Statistics*. New York: McGraw Hill.

Gaito, John. 1980. "Measurement Scales and Statistics: Resurgence of an Old Misconception." *Psychological Bulletin* 87 (3):564–67. https://doi.org/10.1037/0033-2909.87.3.564.

Heilmann, Conrad. 2015. "A New Interpretation of the Representational Theory of Measurement." *Philosophy of Science* 82 (5):787–97.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement Volume I: Additive and Polynomial Representations*. New York: Academic Press.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 2006. *Foundations of Measurement Volume II: Geometrical, Threshold, and Probabilistic Representations*. Mineola, NY: Dover.

Krantz, David H., and Amos Tversky. 2006. *Foundations of Measurement Volume III: Representation, Axiomatization, and Invariance*. Mineola, NY: Dover.

Larroulet Philippi, Cristian. 2021. "On Measurement Scales: Neither Ordinal nor Interval?." *Philosophy of Science* 88 (5):929–39.

[11] In addition to Michell (1986), see Heilmann (2015) and Wolff (2020).

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press

Larroulet Philippi, Cristian. 2022. "Against Prohibition (Or, When Using Ordinal Scales to Compare Groups Is OK)." *British Journal for the Philosophy of Science:721759.* https://doi.org/10.1086/721759.

Lawler, Insa, and Georg Zimmermann. 2021. "Misalignment between Research Hypotheses and Statistical Hypotheses: A Threat to Evidence-Based Medicine? " *Topoi* 40 (2):307–18. https://doi.org/10.1007/s11245-019-09667-0.

Loftus, Geoffrey R. 1978. "On Interpretation of Interactions." *Memory & Cognition* 6 (3):312–19. https://doi.org/10.3758/BF03197461.

Lord, Frederic M. 1953. "On the Statistical Treatment of Football Numbers." *American Psychologist* 8 (12):750–51. https://doi.org/10.1037/h0063675.

Lord, Frederic M. 1954. "Further Comment on 'Football Numbers.'" *American Psychologist* 9 (6):264–65. https://doi.org/10.1037/h0059284.

Marcus-Roberts, Helen M., and Fred S. Roberts. 1987. "Meaningless Statistics." *Journal of Educational Statistics* 12 (4): 383–94. https://doi.org/10.3102/10769986012004383.

Mari, Luca, Mark Wilson, and Andrew Maul. 2023. *Measurement across the Sciences: Developing a Shared Concept System for Measurement*. Cham, Switzerland: Springer Nature.

Michell, Joel. 1986. "Measurement Scales and Statistics: A Clash of Paradigms." *Psychological Bulletin* 100 (3):398–407. https://doi.org/10.1037/0033-2909.100.3.398.

Michell, Joel. 1997. "Quantitative Science and the Definition of *Measurement* in Psychology." *British Journal of Psychology* 88 (3): 355–83. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x.

Narens, Louis. 2012. *Theories of Meaningfulness*. East Sussex, UK: Psychology Press.

Roberts, Fred S. 2009. *Measurement Theory: Volume 7: With Applications to Decisionmaking, Utility, and the Social Sciences*. Reissue ed. Cambridge: Cambridge University Press.

Senders, Virginia L. 1958. *Measurement and Statistics: A Basic Text Emphasizing Behavioral Science Applications*. Oxford: Oxford University Press.

Siegel, Sidney, and N. John Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. Boston, MA: McGraw Hill.

Stevens, Stanley S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684):677–80. https://doi.org/10.1126/science.103.2684.677.

Suppes, Patrick, and Joseph L. Zinnes. 1962. "Basic Measurement Theory." Technical Report 35. Stanford: Stanford University. https://web.stanford.edu/group/csli-suppes/techreports/IMSSS_45.pdf.

Tal, Eran. 2015. "Measurement in Science." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University Press. https://plato.stanford.edu/Entries/measurement-science/.

Tal, Eran. 2021. "Two Myths of Representational Measurement." *Perspectives on Science* 29 (6):701–41.

Thomas, Hoben. 2006. "Measurement Structures and Statistics." In *Encyclopedia of Statistical Sciences*, edited by Samuel Kotz, Campbell B. Read, N. Balakrishnan, and Brani Vidakovic. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471667196.ess1591.pub2.

Velleman, Paul F., and Leland Wilkinson. 1993. "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading." *The American Statistician* 47 (1):65–72. https://doi.org/10.1080/00031305.1993.10475938.

Wagenmakers, Eric-Jan, Angelos-Miltiadis Krypotos, Amy H. Criss, and Geoff Iverson. 2012. "On the Interpretation of Removable Interactions: A Survey of the Field 33 Years after Loftus." *Memory & Cognition* 40 (2):145–60. https://doi.org/10.3758/s13421-011-0158-0.

Wilson, Thomas P. 1971. "Critique of Ordinal Variables." *Social Forces* 49 (3):432–44. https://doi.org/10.1093/sf/49.3.432.

Wolff, Jo E. 2020. *The Metaphysics of Quantities*. Oxford: Oxford University Press. https://books.google.com/books?hl=en&lr=&id=UzjpDwAAQBAJ&oi=fnd&pg=PP1dq=wolff+metaphysics+of+quantities&ots=W1hfpRagFr&sig=hPrUe76Troygwjrd0tPVpSZFG8g.

Cite this article: Mayo-Wilson, Conor. 2025. "Measured Inference: Scales, Statistics, and Scientific Inference." *Philosophy of Science* 92 (5):1205–1214. https://doi.org/10.1017/psa.2025.10173

https://doi.org/10.1017/psa.2025.10173 Published online by Cambridge University Press