psych

MDPI

Article

Scale Type Revisited: Some Misconceptions, Misinterpretations, and Recommendations

Leah Feuerstahler

Department of Psychology, Fordham University, Bronx, NY 10458, USA; lfeuerstahler@fordham.edu

Abstract: Stevens's classification of scales into nominal, ordinal, interval, and ratio types is among the most controversial yet resilient ideas in psychological and educational measurement. In this essay, I challenge the notion that scale type is essential for the development of measures in these fields. I highlight how the concept of scale type, and of interval-level measurement in particular, is variously interpreted by many researchers. These (often unstated) differences in perspectives lead to confusion about what evidence is appropriate to demonstrate interval-level measurement, as well as the implications of scale type for research in practice. I then borrow from contemporary ideas in the philosophy of measurement to demonstrate that scale type can only be established in the context of well-developed theory and through experimentation. I conclude that current notions of scale type are of limited use, and that scale type ought to occupy a lesser role in psychometric discourse and pedagogy.

Keywords: psychometrics; scale type; interval-level measurement

check for updates

Citation: Feuerstahler, L. Scale Type Revisited: Some Misconceptions, Misinterpretations, and Recommendations. *Psych* **2023**, *5*, 234–248. https://doi.org/10.3390/psych5020018

Academic Editors: Alexander Robitzsch and Thomas E. Schläpfer

Received: 9 November 2022
Revised: 22 December 2022
Accepted: 31 March 2023
Published: 4 April 2023

1. Introduction

The emergence of psychology as a distinct area of study in the late nineteenth and early twentieth centuries was accompanied by a heated debate over whether psychology should be considered a truly scientific discipline. At the time, the crux of this debate centered around the possibility of psychological measurement (though this is by no means the full extent of the discussion about whether psychology is a scientific discipline, a discussion which is still ongoing, e.g., [1]). In 1932, a committee of physicists, psychologists, and other academics was formed to debate the use of the terms *measurement* and *quantitative estimates*, and to report on whether quantitative measurement was possible within psychology. The psychological measures that this committee debated were not Likert scales, nor scores on standardized tests. Instead, the committee was primarily concerned with psychophysical measurements, such as those used by such early psychologists as Fechner and Wundt [2]. Eight years later, the committee issued their final report [3] using the sone scale of subjective loudness as their baseline example [4], a psychophysical measure developed by S. S. Stevens. This report detailed how, even after several debates and the formation of subcommittees, there was no consensus as to whether psychology could be a proper scientific discipline. The lack of consensus stemmed from diverging perspectives about whether serious scientific discoveries could be built on measurements for which the appropriateness of concatenating values, taking ratios, etc., had not been proven. In the aftermath of the Ferguson report (and in reaction to his sone scale being singled out as a controversial example), Stevens [5] first proposed the idea that scales could be nominal, ordinal, interval, or rational. One motivating factor for Stevens was to place psychological and educational measurement (PEM) within the same framework as physical measurement; that is, to emphasize that even if there were differences in type, PEM measurement was possible and sensible. One implication of Stevens's taxonomy is that empirical concatenation operations, such as those demanded by some members of the Ferguson committee, are not required for measurement. Instead, Stevens associated different types of measurement scales with different types of

observations, such as determinations of the equality, order, differences, or ratios between observations.

Stevens's classification includes four distinct scale types: Nominal scales are those for which scale values are mere classifications, and for which any one-to-one transformation of those scale values preserves the meaning of the scale. Ordinal scales are those for which scale values reflect the relative ordering of observations, and for which any monotonic transformation preserves the meaning of the scale. Interval scales are those for which equal differences in scale values reflect equal differences in the measured property, and for which linear transformations preserve the meaning of the scale. Ratio scales retain the same properties as interval scales, but with the added condition that the scale has a meaningful zero point such that only multiplicative transformations preserve the meaning of the scale. In the ways in which they are commonly interpreted today, nominal and ratio scales spark little controversy. However, in discussions relating to PEM, scale type is usually brought up when distinguishing between ordinal-level scales and interval-level scales. As such, much of the following discussion will focus on the distinction between ordinal and interval scales.

Although Stevens's classification has numerous examples in the physical sciences, uncontroversial examples are elusive in the non-physical sciences. For ordinal scales, Stevens gives the example of the hardness of minerals (physical), conceding that psychological properties such as intelligence and personality are most likely measured at the ordinal level. For interval scales, he gives the examples of Centigrade and Fahrenheit temperature scales, and time (such as calendar dates). He is less specific regarding the extent to which interval-level measurement can be applied to psychological properties. Stevens writes, "[m]ost psychological measurement aspires to create interval scales, and it sometimes succeeds. The problem usually is to devise operations for equalizing the units of the scales—a problem not always easy of solution but one for which there are several possible modes of attack" [5] (p. 679). Unfortunately, Stevens provides no example of when interval-level measurement has succeeded (though he argues that sones are measured as a ratio-level scale), nor what modes of attack might be appropriate for measuring constructs like intelligence and personality. Instead, he states that "In most cases a formulation of the rules of assignment discloses directly the kind of measurement and hence the kind of scale involved. If there remains any ambiguity, we may seek the final and definitive answer in the mathematical group-structure of the scale form..." (p. 680). Although this may provide sufficient guidance for measures like the sone scale, which are psychophysical and for which the construct may be experimentally manipulated, the implications for psychological assessment are much less clear.

From the ambiguity present in Stevens's original work arose a new field of mathematics; namely, representational measurement theory (RMT) [6], based on the idea that there should be a direct correspondence between empirical relations among objects and numerical relations among measurements of those objects. The field of RMT can be understood as a formalization of Stevens's scale types; that is, a set of testable conditions that must hold for each scale type. One compelling feature of RMT is that it can be applied to both extensive quantities (i.e., those for which we can empirically concatenate objects, such as mass and distance) and intensive quantities (i.e., those for which it is impossible to concatenate objects to produce a sum of their individual measurements, such as temperature and momentum). The ability to test the structure of intensive quantitative (i.e., interval- or ratio-level) variables can theoretically be applied to any PEM measure, and I will go on to discuss some of the applications and challenges of this approach later in this paper. Although there are a few notable measures that derive from the RMT models, particularly the additive conjoint model [7,8], the richness of RMT has had relatively little impact on PEM practices.

Whether in the form of RMT or not, Stevens's scale types have had a profound impact on the practice of psychological and educational research. Perhaps in part due to Stevens's work, the field of psychology has changed dramatically from the discipline to which the Ferguson report (and Stevens, by extension) first reacted [9]. Nevertheless, Stevens's work

gave psychologists reason to believe that with enough effort, measurement was not only possible but achievable at an interval level. Because Stevens provided little guidance as to how to verify scale type for the types of measurements that were being increasingly used for PEM (e.g., standardized educational tests, Likert scales, symptom inventories, etc.), researchers (especially those working prior to the development of RMT) were left to their own devices, applying Stevens's ideas as they saw fit. In the following sections, I outline several instances in which scale type has since been invoked in conflicting or contradictory ways. I demonstrate that the nature, implications, and verification of interval-level PEM scales are often poorly understood, arguing that an emphasis on scale type leads to overconfidence in newly developed measures and stifle scientific progress. Many of the ideas that I raise and discuss below draw heavily on previously published works that attempt to clarify these issues. However, the persistence of these poorly understood ideas surrounding scale type warrants emphasis and re-examination.

2. The Meaning of Interval-Level Measurement Is Poorly Understood

There are at least three distinct interpretations of Stevens's scale types: the representational approach, the operational approach, and the classical approach [10]. From the representational perspective, the assignment of numbers to observations should reflect empirical relations among observations, such as equivalence, ordering of values, ordering of differences, and concatenation relations (which correspond to nominal, ordinal, interval, and ratio measurement levels, respectively). For example, it is possible to concatenate two identical rulers by laying them end-to-end, such that the length spanned by the two rulers is twice the length of either. Therefore, we may assign numerical values to lengths as ratios of either ruler length and consider the resulting values to be ratio-level measurements. From the representational perspective, determining levels of measurement for many physical variables (e.g., momentum, which is derived rather than extensive) and most psychological variables is not so simple (and is likewise subject to misunderstanding, as described in a later section). However, the models of RMT are able to characterize the relationships between derived measures of different scale types, as these models delineate the types of empirical relationship that should hold for each type of measurement scale; see [11–13] for further discussion and critique of the applicability of RMT in psychometrics. From the operational perspective, measurement concerns the use of consistent rules to assign numbers to observations. Here, scale type is a matter of the operations for which the test will be used; for example, a researcher may state that responses to a 4-point Likert scale constitute an ordinal-level measurement because, while the numerical values are assigned in a consistent way, we cannot be confident that the intervals between scale points are equal. In its purest form, operationalism is not concerned with the correspondence between the numerical assignments and the underlying phenomenon; instead, scale type is relevant only in that mathematical operations performed upon the measurements should correspond to the stated scale type. The classical perspective differs from representationalism and operationalism in that measurement relations are discovered and not assigned; from the classical perspective, only quantitative attributes (i.e., attributes at the interval or ratio level) can be measured, and it is the task of the scientist to demonstrate empirical evidence for a variable's quantifiability. Michell [10] claims that although the classical perspective was once the common mode of scientific measurement, it has largely been replaced by representationalism and operationalism.

The operational, representational, and classical perspectives differ in a number of ways, including in terms of the steps involved for measurement within a given scale type. From the operationalist perspective, scale type concerns the techniques used to construct the scale, as well as the score manipulations that will be used. From the representationalist perspective, scale type concerns the demonstrated mathematical properties of the scale. In fact, both of these perspectives on scale types can be found in the 1946 Stevens paper: "These classes are determined both by the empirical operations invoked in the process of 'measuring' [operationalist] and by the formal (mathematical) properties of the scales

[representationalist]. Further—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to the empirical data depend upon the type of scale against which the data are ordered [operationalist]". Finally, from the classical perspective, scale type constitutes the state of knowledge about the underlying phenomenon in question. One major way in which the classical perspective differs from the operationalist and representationalist perspectives is that the scale scores are discovered rather than assigned; in both the operationalist and representationalist approaches, determinations of scale type are decided prior to empirical research. It is therefore important to settle the issue of scale type before conducting further research. From the classical perspective, scale type does not need to be settled prior to empirical research, instead the nature of the scale is a valid and necessary topic for empirical research.

In practice, it is rare to find these perspectives stated in clear terms, and there appears to be a disconnect between these abstract philosophical perspectives and their application. For example, although the operationalist and representationalist perspectives are sometimes characterized as attempts to avoid realism [14], realism is arguably the most coherent way to interpret such psychometric models as latent variable models [15]. Similarly, realism is difficult to avoid in practice [14], and more recent work has discussed definitions of representationalism [16] and operationalism [17,18] that do not preclude realism. Although it has been argued that these realist interpretations are based on logical errors [19], in practice the most prevalent perspectives on scale type combine the ideas of either operationalism or representationalism with certain elements of the classical perspective. I will broadly characterize these hybrid perspectives as being "operational-ish" and "representational-ish", since they do not represent the purest form of any philosophical perspective; while the operational-ish perspective tends to use the number-assigning tools of operationalist theory and the representational-ish perspective tends to use the number-assigning tools of representationalist theory, both incorporate realist interpretations. To draw this distinction, I will refer to broad trends in methods and interpretations that I have found useful to explore. Of course, there remains considerable heterogeneity within each perspective, and many researchers maintain perspectives that do not fit neatly into either.

In practice, most PEM researchers agree that measuring something means assigning numbers to observations in a systematic way. The framing of this idea, though standard in many textbooks, necessitates either the representationalist or operationalist perspective and excludes the classical perspective (which involves number assigning, certainly, but is fundamentally about discovery). However, from a representational-ish or operational-ish view, researchers are compelled at one end to make assignments that fit into a specific scale type, but later to discuss these measurements as if they are real discovered quantities. Because of the confusion of philosophical viewpoints, there is subsequent confusion about the proofs for, and implications of, levels of measurement (as will be discussed in later sections). If measures are assigned, then the measurement scales must be determined prior to research. If they are discovered, then which scale units are most appropriate is an empirical question.

3. Appropriate Evidence for Interval-Level Measurement Is Unclear

Following the distinct notions of interval-level measurement described in the previous section, there are conflicting ideas as to what evidence is sufficient to demonstrate that a particular scale constitutes a certain level of measurement. Many PEM researchers will acknowledge that many PEM scales are at the ordinal (rather than the interval) level of measurement, yet interval-level measurement is often perceived as an achievable goal (e.g., some studies purport to transform ordinal-level scales to interval-level scales [20–22]). One problem is that researchers who use the operational-ish or representational-ish perspectives have different ideas of what types of evidence are appropriate or needed to justify their claims. Please note that the ideas that I discuss in the following paragraphs are not always discussed in clear terms, and individual authors may disagree with my characterizations of

their work. My goal in providing examples is not to single out individual studies but to exemplify those patterns of thought and interpretation that appear more broadly.

In the operational-ish view, many believe that interval-level PEM scales are an ideal from which observed measurements fall short. These shortcomings are believed to be a matter of degree rather than type. From this perspective comes the idea that interval-level scales are more accurate than ordinal-level ones [23], and the two commonly cited ways in which PEM scales fail to be truly interval-level are that they are plagued by measurement errors, and that they are discrete rather than continuous [24] (cf. the five types of ordinal variables described by [25]). From this perspective, one common idea is that continuity (a term that is interpreted variously but is often used to refer to scales that have many unique points) implies interval-level measurement. This stems from the misunderstanding that ordinal scales must be discrete and interval scales must be continuous. In the psychometrics literature, this idea surfaces in recommendations to replace Likert scales with a continuous line [26], or to use a large number of Likert scale categories [27]. Even though these may be useful strategies to design a system of measurement, they do not necessarily bring the scale units closer to an interval level. It is true that interval-level scales are theoretically continuous, but any realized measurements are necessarily discrete, due (at least in part) to the measurement errors induced by rounding. The requisite condition of continuity for interval-level scales comes from mathematical real analysis, rather than the folk meaning of continuous as "having many unique points". It is easy to demonstrate that the number of scale points is largely irrelevant to the scale type; for example, the distance of world rivers in kilometers (a ratio-level variable) may be nonlinearly transformed into log-kilometers (which might be considered an ordinal-level variable, since this transformation does not preserve ratio-level invariance), yet has just as many unique points, and thus bear just as much of a surface-level appearance of continuity.

Another way in which continuity is confused for interval-level measurement is in the use of latent variable models such as factor analysis, item response theory, and structural equation models. Here, I will focus on this idea in the context of reflective item response models (IRMs), where claims that the latent variable is measured at the interval level measurement are common [28–31]. Parametric IRMs such as the two- and three-parameter logistic models characterize the probability of responding in two or more categories as a function of a latent variable. By observing a number of responses that are all determined by the same latent variable (or construct), an individual's position along that latent variable can be estimated. Once the parametric form of an item response model is chosen, the latent variable scale is determined up to a linear transformation. From this fact, it has been argued that parametric IRMs lead to interval-level measurement and that, in demonstrating that the data fit a parametric IRM, one can claim that their scale is interval-level. There are several problems with this logic. First, the parametric form of the item response model is chosen by the researcher [32]; unless an argument is made that only the chosen parametric form should be used to characterize the relationship between response probabilities and the latent variable (as is argued with regard to the Rasch model, discussed next), there is little justification to claim that IRMs inherently lead to interval-level measurement. It can be demonstrated [33] that for any given IRM, any monotonic transformation of the latent variable—linear or nonlinear—results in another IRM that fits equally well and makes identical predictions about response behavior. In addition, it has been shown that IRT model misspecification can result in nonlinear distortions of the metric [34,35]. Moreover, applying fit statistics is a poor way to demonstrate that data were generated from the chosen model. Fit statistics simply tell how well the chosen model accounts for the data. Instead, to make claims about the data-generating process for item response data, other models must be ruled out. This information is not given by fit statistics (nor were fit statistics developed to address this concern). Considering Lord's [33] result regarding nonlinear scale transformations, the claim that the latent variable metric of item response theory is interval-level is unfounded.

A similar story can be told about Thurstone's method of paired comparisons. This method has been argued to provide equal-interval scales by a number of applied (e.g., [36,37]) and technical (e.g., [38] p. 125, [39] pp. 57–58) authors. However, this conclusion rests on the assumption that comparative judgments are normally distributed [38]. Similar to the assumption of the function form of the item response model, different choices of the distribution of judgments will lead to non-monotonically related scales. Although these assumptions are convenient for model fitting and estimation, they are typically interpreted outside any effort to establish their correspondence to real phenomena.

From the representational-ish perspective, certain models must be used in order to claim interval-level measurement. Perhaps the most common implementation of this idea appears when researchers claim that fitting the Rasch model yields interval-level measures of persons and items. This idea stems from the close correspondence between the one-parameter Rasch model and the additive conjoint model (ACM) of RMT [8]. The ACM [40] establishes interval-level measurement for a set of three variables by showing that (a function of) the additive relationship between two of those variables implies a certain value for the third variable. When applied to the Rasch model, the logistic function of the difference between person scores and item scores implies particular response probabilities. In this way, the Rasch model does follow the form of the ACM. Many applied authors have taken this result to mean that fitting the Rasch model to their data is sufficient to achieve interval-level measurement (e.g., [20,22,41]). Others have taken these results further by arguing that the Rasch model is not only sufficient, but necessary for interval-level measurement [21,42]. In each of these papers, adequate fit statistics (e.g., "acceptable" values of infit and outfit) were the primary criteria for asserting that their scales were interval-level. One problem with this approach is that fit statistics designed to assess the correspondence between data and the Rasch model are not adequate for establishing the fit of the data to the ACM. Several authors (e.g., [43,44]) have noted that the fit statistics usually used for the Rasch model do not formally test the axioms of additive conjoint measurement. Instead, other tests of the ACM axioms have been developed [45,46], but have not yet been widely used.

Although there is a formal correspondence between the Rasch model and the ACM, there are also important practical differences; for one, none of the quantities which are purported to be interval-level in the Rasch model are directly observable. More importantly, item response probabilities are not only unobserved but unobservable. This simple fact poses several problems for the relationship between the Rasch model and the ACM. First, because the observed data are discrete scores rather than probabilities, one cannot directly test the axioms of additive conjoint measurement. At best, one can test whether or not the data appear to fit the model. Doing so requires a theory of the residual structure of the data, and the ACM suggests no theory of how residuals ought to be structured. Therefore, by formulating a probabilistic version of the additive conjoint model, one must make assumptions external to RMT. A second problem with this idea that the Rasch model provides interval-level measurement is that the logistic function which relates differences in person (θ) and item (b) scores to response probabilities is not the only function that can serve this role. In theory, any continuous cumulative distribution function of ($\theta - b$), such as the complementary log–log link [47], satisfies the axioms of ACM in the same way that the Rasch model does [48] (p. 249). Models such as the complementary log–log item response model will lead to scales that are nonlinearly related to the Rasch scale, throwing into serious question the adequacy of probabilistic formulations of the ACM for producing interval-level measurement. To be clear, the special advantage of the logistic function in Rasch measurement is that it uniquely provides sufficient statistics for both the item and person parameters. This property is known as specific objectivity and is indeed remarkable and useful, but it is neither required by the axioms of additive conjoint measurement nor does it appear to have any special role within any RMT framework.

Finally, researchers who continue to insist that the Rasch model implies interval-level measurements should bear in mind that the ACM is only one way in which interval-

level measures can be constructed within RMT. RMT comprises models that can include more complex relationships among variables than the ACM and still yield interval-level measurement [49]. It has even been suggested that the probabilistic forms of these more complex conjoint structures would take the form of non-Rasch item response models, such as the two-parameter model [50] (p. 361). In summary, the Rasch model is not sufficient for interval-level measurement because it is an inexact (i.e., probabilistic) realization of the additive conjoint model, and it is not a necessary condition for interval-level measurement because alternative models to the ACM can also lead to interval-level measurements. The many advantages of the Rasch approach notwithstanding, a full understanding of the literature casts serious doubt upon the strict position that Rasch measurement implies interval-level measurement.

In all of the models discussed in this section, the claims made about interval-level measurement would be inconsequential if not for the tendency of researchers to interpret interval-level-ness as a real and newly discovered feature of their scale. As discussed earlier, neither the representationalist nor operationalist perspectives necessarily entail realism, nor can they provide any proof for or against the real existence of the measured quantity. Rather than advance theory, such interpretations may serve instead to stifle further inquiry into the most meaningful or interpretable units of the scale. As I will argue next, the contribution to theory to the establishment of useful and reliable score intervals is a scientific question that ought not to be resolved quickly. I do not mean to suggest that the scoring scheme provided by any of these models is undesirable or incorrect. At their most basic level, different methods provide different strategies for assigning scores that may or may not be useful or meaningful.

4. The Implications of Interval-Level Measurement Are Poorly Understood

Many PEM researchers strive for interval-level measurement because they believe that interval-level scales are necessary for the statistical tests they wish to conduct on their data. This aspect of Stevens's scale types has received more attention than any other. The idea that different statistical tests are only appropriate for certain types of scales traces back to Stevens's original writing [5]. For example, Stevens lists medians and percentiles as permissible statistics for ordinal-level scales, and the mean, standard deviation, and product-moment correlation as permissible statistics for interval-level scales. As a result, there is a widespread belief that scale type determines the statistical tests that either can or should be run on various measured variables. Stevens seems to imply that parametric tests—including linear regression, *t*-tests, ANOVA, etc.—should only be conducted with interval- or ratio-level variables, and that nonparametric tests should be conducted with ordinal-level variables. These ideas stem from the notion that each scale type is associated with a transformation that preserves score meaning. For example, ratio-level scales are only multiplied by a constant (such as when converting inches to centimeters), while interval-level scales may be transformed by a multiplicative constant and an additive term (such as when converting degrees Fahrenheit to degrees Celsius). So-called "permissible statistics" are those that will yield the same result (e.g., *p*-value, effect size) for any invariant scale transformation.

The "permissible statistics" controversy has raged for the better part of a century, with several methodological and applied researchers weighing in [24,51–59]. The following comments are not meant to be a comprehensive or conclusive response to this controversy, and the interested reader may refer to these writings for a more thorough exploration of these ideas. However, some points must be addressed, as this is perhaps the most common misconception regarding scale type. It is important to note that Michell [10] argues that only the classical (not the operationalist or representationalist) perspective has any implications for the use of statistical procedures. The widespread belief that there is at least some relationship between scale type and statistics suggests that my distinction between operational-ish and representational-ish perspectives (both of which incorporate realist interpretations) more accurately represents actual practice than Michell's classification.

Much of the controversy over permissible statistics emerges from the operational-ish, rather than the representational-ish, perspective. Those following the strictures of RMT feel confident in their use of "interval-level statistics" because they believe they have taken precautions to ensure that their scales actually meet this criterion. When those in the representational-ish camp speak on the issue of permissible statistics, it is typically to motivate why it is important for them to achieve interval-level measurement (i.e., so that they are able to conduct certain analyses) [20,22,41,42]. Although Michell [10] argued that the representationalist perspective does not entail any implications about permissible statistics, I believe that an unspoken realist assumption within the representational-ish perspective explains the persistent belief that there exists a necessary relationship between scale type and statistics. Without the realist belief that the numerical representation reflects the true underlying nature of a variable, the restrictions of permissible statistics would be unnecessary.

Competing ideas about the implications of interval-level measurement thus emerge from within the operational-ish perspective. On one hand is the idea that one should commit to a level of measurement and act according to the strict rule that parametric statistics must only be used with interval- or ratio-level scale and nonparametric statistics must be used for ordinal-level scales [52,60]. A softer version of this perspective [61] argues that, although parametric statistics are not strictly appropriate for ordinal-level measures, it may be beneficial to conduct both the parametric and nonparametric versions of a test, such as the independent samples *t*-test and the Mann–Whitney U test; if the tests agree (presumably on the binary decision of whether to reject the null hypothesis), the parametric results may be reported, but if the tests disagree, the nonparametric test should be reported. This is dubious advice because this procedure guarantees the inflation of the type I error rates and could be easily misinterpreted as providing confirmatory evidence. As such, the cautious applied researcher may believe that they should limit themselves to the tools of nonparametric statistics.

Another perspective is that the meaningfulness of interpretations matters, not levels of measurement. From this perspective, the conversation should not be about what statistics are permissible, but about what statistics lead to meaningful conclusions [54]. Borrowing from a classic example, it is difficult to imagine that any interesting statistical results would occur from parametric statistical analyses of football numbers. However, as Lord [53] humorously illustrated, there is nothing inherent about football numbers that require them to be used only in statistical tests designed for nominal data. Scale type can be context-dependent; in Lord's story, sophomore football players tampered with the dissemination of football numbers so that freshmen players had lower numbers. If parametric statistics are used to evaluate football numbers, interval meaning is ascribed to the numbers [57] (see [62]). Whether it is meaningful or interesting to evaluate football numbers as an interval-level variable cannot be satisfactorily resolved by ascribing a scale type to the variable "football numbers" without considering the context and potential use. In the context of designating different players, there is no ordinal meaning to football numbers, but in the context of social status, there may be an ordinal- or interval-level meaning ascribed. The argument that measurement scales should be meaningful often entails notions of realism. However, "meaningfulness" is used descriptively in these articles, and guidance for how to demonstrate or effectively argue for meaningfulness is lacking.

Before moving on from the topic of permissible statistics, there are some scientific questions for which the equal-interval property of a scale is of critical importance. These questions concern the analysis of differences between scale values, such as when measuring change over time or in value-added assessment. As the above discussion demonstrates, researchers from different philosophical perspectives will accept different forms of evidence, depending on whether they ascribe to the representational-ish or operational-ish perspectives. Although this may seem to lead to an impasse, the fact that both perspectives desire to interpret their results realistically suggests that there is room for common ground. Specifically, I believe that a more experimental approach to determining scale intervals (an

idea borrowed from the classical perspective) may provide a path forward. An example of this experimental approach is Stevens's sone scale [4] (the same scale that was used as a controversial example by the Ferguson committee). The sone scale is a measure of perceived loudness that was established using multiple experimental methods. These methods included asking experimental subjects to vary the sound intensity of tones until one tone was judged to be half as loud as another, presenting two tones of equal loudness at once, and then asking them to compare that to a third tone. Based on the results of these experiments, Stevens concluded that the sones have a ratio level of measurement [5], even though they are nonlinearly related to other (objective) measures of loudness. The example of the sone scale has a somewhat limited application in PEM research where empirical concatenation operations are not possible and score ratios are of lesser interest. However, it does illustrate the possibility of establishing meaningful intervals through experimentation, and later I will highlight some modern movements in this same direction.

5. Modern Philosophical Perspectives

In both the operational and representational interpretations of scale type, measurements are logically situated prior to theory and prior to applied scientific research. From this perspective, the mapping of measurements onto the empirical world is wholly abstracted from theoretical considerations. However, modern philosophical perspectives on measurement argue that all measurement (not just PEM measurement) is necessarily and inextricably theory-laden [63]. Writings in this vein take particular issue with the abstract nature of RMT: "RMT's role with respect to measurement theory is therefore akin to that of axiomatic probability theory with respect to quantum mechanics: both accounts supply rigorous analyses of indispensable concepts (scale, probability) but not the conditions of their empirical application" ([64] pp. 77–78). In addition, identifying a physical correlate to which measurements correspond is not a full account of measurement. In the modern view, measurements are designed to answer specific questions, and the answers provided should be meaningful and relevant with respect to theory [65]. As the following example illustrates, the modern perspective argues that theory-driven measurement—which guides theory-based experimentation—is key to scientific progress.

The modern philosophical approach to measurement can be illuminated by an understanding of the history of temperature measurement. Temperature is a compelling case study because (a) there is consensus that this variable is measured on an interval-level scale—in fact, it is the primary example that Stevens (1946) gives for interval-level scales (Temperature in degrees Kelvin is a ratio-level variable. Temperature in degrees Celsius or Fahrenheit, being an affine transformation of temperature, is considered to be at the interval level); (b) its establishment as an interval-level measurement variable occurred fairly recently (i.e., in the nineteenth century) and is well-documented; and (c) it has been used as a case study by several philosophers of measurement [66–68]. Of course, ideas about temperature existed long before its measurement was standardized in the nineteenth century. In fact, thermoscopes are known to have been used around 1600 by Galileo and his contemporaries, and these instruments were based on little more than the empirical phenomenon that gases tend to expand with heat [67]. However, before temperature was definitively established as an interval-level measure, careful work was conducted by Regnault to make extremely precise temperature measurements [66]. Notably, Regnault's approach was atheoretical; in psychometric terms, his measurements were highly reliable, but he showed little concern for validity or theory. None of Regnault's work illuminated the nature of temperature, and therefore when tested in novel situations, the highly reliable thermometers made unrealistic and inaccurate readings. Instead, it was through the experimental work of Black that deterministic relationships were established between measured temperature and the volume of mercury [68]. It was through this work that temperature was established as an interval-level measure. Specifically, Black used experimentation, including combining fluids with different volumes and different temperatures, to understand the deterministic nature of how temperature interacts with other physical attributes such

as volume and substance. It was through this and other careful experimental work that temperature measurement was internationally standardized. Today, the Kelvin scale is not only widely recognized to be a ratio-level scale, but it is a base unit in the International System of Units (SI) and defined precisely in terms of the kinetic energy of particles through the Boltzmann constant [69]. Notably, this current definition was not adopted into the SI system until 2019, far after many properties of temperature were well understood by the scientific community.

There are several lessons that psychometricians can glean from the history of temperature measurement. First, following an argument made by Bringmann and Eronen [66], robustness and validity are far more important than scale type when establishing new measurements and new theories. That is, robust theory can be built on "bad" measurements. Second, it is not necessary to first establish scale type before "treating the scale as the desired type". As Black's experiments show, the experimental manipulations not only occurred prior to, but were vital to, proving the interval nature of temperature measurements. Finally, psychometricians should take this case study as a sobering example. For many psychological variables, it is hard to imagine the experimental operations that would need to take place to definitively establish interval-level measures from a classical perspective. Even though needed methods for PEM measurement are in a primitive state, this does not preclude experimentation with (and refinement of) existing measures. Indeed, it may only be through the application of these measures that the PEM sciences can progress to a point where scale type is relevant.

6. Discussion

Earlier, I described the representationalist, operationalist, and classical perspectives of measurement and described how these do not appear in their pure forms in the minds of many researchers. Although I maintain that there are aspects of the classical perspective that warrant further attention, I have stopped short of claiming that this perspective is preferable and that the others ought to be abandoned. Instead, my aim has been to identify the poor reasoning that stems from each tradition and highlight ways in which greater awareness of the classical perspective may improve practice. Insofar as realist interpretations will be made, it may be useful to assume that measurements are discovered rather than assigned. Stating this assumption explicitly avoids the "pathology" of blindly inferring that measurements reflect an independently existing attribute [70], and I believe that such reasoning is not necessarily a logical error. For instance, Vessonen [17] describes a type of cautious operationalism that allows for a realistic perspective while suspending judgment about the epistemic status of a particular measurement. Realistic interpretations of measurements are both useful and tenuous, and by explicitly acknowledging these points, scientific progress may be made without requiring many of the legalistic practices described earlier in this paper.

6.1. Recommendation 1: Establish Scale Intervals through Experimentation

I believe that the current modes in which most PEM scales are established do not address the question of how to best select the intervals for that scale. This is partly because different researchers have different understandings of what scale type means (and therefore may rely on the scale produced by their chosen model), but also because our current measurement tools are often crude, with many different measures available for apparently similar constructs (e.g., [71,72]). Combined with the tenuous nature of claiming that there exists a quantitative underlying phenomenon to measure, I believe that PEM measurement would benefit from a more honest reckoning with these facts. When we (skeptically) treat appropriate scale intervals as quantities to be discovered through experimentation, we are not required or expected to determine scale type before empirical research. In this mode, the scale intervals we propose can be justified through evidence-based arguments rather than on variously understood notions of scale type or modeling assumptions (cf. the pragmatic applied numerics approach suggested by Barrett [73]).

Creative experimental methods (possibly inspired by the classical perspective) may be far more useful for establishing meaningful and justifiable scales and scale intervals than routine model fitting. There are some examples of this strategy in the literature [74,75], including Stevens's sone scale [4]. One strategy to deal with the scale dependency of these results could be to anchor the scale to external variables or clinically meaningful outcomes [75]. Blanton and Jaccard [75] have discussed various challenges to this approach and suggest a strategy of building consensus among researchers as to what contextually meaningful scale points may be. Other examples of scales that are anchored to external variables include sum-score-like measures [76] or proportions-of-domain. The literature on clinically meaningful and minimally important difference scores [77–79] is also in the spirit of this recommendation, as it benchmarks scores with meaningful outcomes (though this literature does not always consider alternate units for measurements). A number of methods exist for transforming scores to different metrics [80], and these can be adapted to use external variables to help identify meaningful metrics. Methods such as those developed by Ramsay and Wiberg [76], as well as by Feuerstahler [81], begin to address the technical challenges involved in applying model-based psychometrics to user-defined scaling metrics.

This experimental approach to scale scores does not mean that we can neither use nor trust our intervals derived from the original measurements. In fact, this may be a logical starting point. However, acknowledging that the readily available scale units may not be the most useful representation can lead to fruitful research into scale intervals. If a researcher finds comfort in using RMT-derived methods to establish their units, that is not objectionable per se, but they should also be aware that these are not unassailable units. Even for research topics where interval-level measurements appear to be essential (e.g., measurement of growth, and value-added assessment), efforts to establish those scales prior to experimentation may be fruitless or instill the researcher with a false sense of confidence. It is not necessarily the case that questions of growth and added value should be set aside until interval-level scales are established. Instead, researchers should be aware that their results are scale-dependent, and that at the current state of scientific knowledge there may be no way to resolve the scale dependency of these results. This should not stifle research into these sorts of questions. Rather, these sorts of questions might be critical to simultaneously refining both the measuring instruments and theory; in other words, numerical assignment and discovery of scale properties can be in continual dialogue, with empirical research informing both.

Finally, it should be acknowledged that the role of scale type (the focus of this paper) is not the only consideration for improving the practice of number-assigning in PEM research. Other ideas, such as focusing on traceability among measurements [82,83] and the possibility of non-quantitative data structures [73], should also be part of the ongoing dialogue between applied researchers, psychometricians, and philosophers of science.

6.2. Recommendation 2: Reconsider How We Teach Levels of Measurement

Stevens's scale types appear in the vast majority of introductory textbooks on psychology, and statistics for the social sciences, among many more. Typically, descriptions of these scale types are presented uncritically, which reflects the impact of Stevens's work but does not reflect the subsequent controversy. Ultimately, Stevens's work is an attempt to synthesize various ideas about measurement and quantity and to "widen the playing field" to include measurements of a discrete and/or ordinal nature. His classification is a proposed framework, not a law of nature discovered by Stevens and supported by empirical evidence. Although many researchers through the years have found this distinction useful, it has also caused confusion and misunderstanding as much as it has enlightened the field of psychometrics. For practical purposes, I find it useful for introductory courses to replace pedagogy on scale type with (a) a distinction between discrete and continuous quantities, and (b) an emphasis on the meaning and interpretation of measurements; i.e., that there is always a layer of inference from our measures to any underlying phenomena.

The instructor may also emphasize that the distinction in (a) is not a strict one. Count and Likert scale data could be analyzed as continuous quantities or could be analyzed discretely using Poisson, negative binomial, or ordered logistic models. In more advanced classes, it could be useful for students to learn to treat their units with skepticism and to develop research strategies for experimentally discovering meaningful units. Such changes in pedagogy would more honestly reflect statistical practice, and would encourage students to question, manipulate, and deeply understand their scale units and the processes from which they arise.

7. Conclusions

For most of a century, the concepts of nominal, ordinal, interval, and ratio-level scales have permeated PEM research. Although many find this classification to be a useful way to describe different types of scales, I have argued in this paper that these ideas have led to many confusing and conflicting interpretations. On one hand, overly cautious researchers feel compelled to limit the statistical tools they use (and thereby the questions they can ask of data). On the other hand, those determined to achieve interval-level measurement, upon fitting a model promising as much, may become overly confident in the universality of their units.

Of course, the fact that a concept like "interval-level measurement" is variously interpreted is not reason enough to abandon it. However, I do believe that the time has come for PEM researchers to seriously reconsider the benefits and harms of continuing to perpetuate levels of measurement as an important aspect of psychometric scales. I do not advocate for the total abandonment of Stevens's work, RMT, nor any other fruitful line of research. However, I do believe that it is time to make scale type a smaller part of the conversation on PEM measures and measurement quality, and where it continues to be invoked, to discuss these concepts in clearer and more consistent terms.

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: Many thanks to [blinded], who provided valuable feedback on an earlier version of this manuscript. All remaining errors and opinions are the author's.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACM	Additive Conjoint Model
IRM	Item Response Model
PEM	Psychological and Educational Measurement
RMT	Representational Measurement Theory

References

1. Trendler, G. Measurement theory, psychology and the revolution that cannot happen. *Theory Psychol.* **2009**, *19*, 579–599. [CrossRef]
2. Jones, F.N. History of psychophysics and judgment. *Handb. Percept.* **1974**, *2*, 1–22.
3. Ferguson, A.; Myers, C.; Bartlett, R.; Banister, H.; Bartlett, F.; Brown, W.; Tucker, W. Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Rep. Br. Assoc. Adv. Sci.* **1940**, *2*, 331–349.
4. Stevens, S.S. A scale for the measurement of a psychological magnitude: Loudness. *Psychol. Rev.* **1936**, *43*, 405. [CrossRef]
5. Stevens, S.S. On the theory of scales of measurement. *Science* **1946**, *103*, 677–680. [CrossRef]
6. Suppes, P.; Zinnes, J.L. Basic measurement theory. In *Handbook of Mathematical Psychology*; Luce, R.D., Bush, R.R., Galanter, E., Eds.; John Wiley and Sons: New York, NY, USA, 1961; Volume I.
7. Kahneman, D.; Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **1979**, *47*, 263–292. [CrossRef]
8. Perline, R.; Wright, B.D.; Wainer, H. The Rasch model as additive conjoint measurement. *Appl. Psychol. Meas.* **1979**, *3*, 237–255. [CrossRef]

9. Danziger, K. *Constructing the Subject: Historical Origins of Psychological Research*; Cambridge University Press: Cambridge, UK, 1994.
10. Michell, J. Measurement scales and statistics: A clash of paradigms. *Psychol. Bull.* **1986**, *100*, 398. [CrossRef]
11. Michell, J. Representational measurement theory: Is its number up? *Theory Psychol.* **2021**, *31*, 3–23. [CrossRef]
12. Trendler, G. Conjoint measurement undone. *Theory Psychol.* **2019**, *29*, 100–128. [CrossRef]
13. Trendler, G. The incoherence of Rasch measurement: A critical comparison between measurement in psychology and physics. *Personal. Individ. Differ.* **2022**, *189*, 111408. [CrossRef]
14. Borsboom, D.; Cramer, A.O.; Kievit, R.A.; Scholten, A.Z.; Franić, S. The end of construct validity. In *Concept of Validity: Revisions, New Directions and Applications*; IAP Information Age Publishing: Charlotte, NC, USA, 2009.
15. Borsboom, D.; Mellenbergh, G.J.; Van Heerden, J. The theoretical status of latent variables. *Psychol. Rev.* **2003**, *110*, 203–219. [CrossRef]
16. Vessonen, E. Representation in measurement. *Eur. J. Philos. Sci.* **2021**, *11*, 1–23. [CrossRef]
17. Vessonen, E. Operationalism and realism in psychometrics. *Philos. Compass* **2019**, *14*, e12624. [CrossRef]
18. Vessonen, E. Respectful operationalism. *Theory Psychol.* **2021**, *31*, 84–105. [CrossRef]
19. Uher, J. Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* **2021**, *41*, 58. [CrossRef]
20. Hamilton, C.B.; Chesworth, B.M. A Rasch-validated version of the upper extremity functional index for interval-level measurement of upper extremity function. *Phys. Ther.* **2013**, *93*, 1507–1519. [CrossRef]
21. Tennant, A.; Conaghan, P.G. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res.* **2007**, *57*, 1358–1362. [CrossRef]
22. Walton, D.M.; Wideman, T.H.; Sullivan, M.J. A Rasch analysis of the pain catastrophizing scale supports its use as an interval-level measure. *Clin. J. Pain* **2013**, *29*, 499–506. [CrossRef]
23. Reid, J.; Nolan, A.; Hughes, J.; Lascelles, D.; Pawson, P.; Scott, E. Development of the short-form Glasgow composite measure pain scale (CMPS-SF) and derivation of an analgesic intervention score. *Anim. Welf.* **2007**, *16*, 97–104. [CrossRef]
24. Borgatta, E.F.; Bohrnstedt, G.W. Level of measurement: Once over again. *Sociol. Methods Res.* **1980**, *9*, 147–160. [CrossRef]
25. Kampen, J.; Swyngedouw, M. The ordinal controversy revisited. *Qual. Quant.* **2000**, *34*, 87–102. [CrossRef]
26. Allen, I.E.; Seaman, C.A. Likert scales and data analyses. *Qual. Prog.* **2007**, *40*, 64–65.
27. Leung, S.-O. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-Point Likert Scales. *J. Soc. Serv. Res.* **2011**, *37*, 412–421. [CrossRef]
28. Harwell, M.R.; Gatti, G.G. Rescaling ordinal data to interval data in educational research. *Rev. Educ. Res.* **2001**, *71*, 105–131. [CrossRef]
29. Kirisci, L.; Tarter, R.E.; Vanyukov, M.; Martin, C.; Mezzich, A.; Brown, S. Application of item response theory to quantify substance use disorder severity. *Addict. Behav.* **2006**, *31*, 1035–1049. [CrossRef] [PubMed]
30. Mungas, D.; Reed, B.R. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Stat. Med.* **2000**, *19*, 1631–1644. [CrossRef]
31. Sijtsma, K.; Molenaar, I.W. *Introduction to Nonparametric Item Response Theory*; Sage: Thousand Oaks, CA, USA, 2002; Volume 5.
32. Yen, W.M. The choice of scale for educational measurement: An IRT perspective. *J. Educ. Meas.* **1986**, *23*, 299–325. [CrossRef]
33. Lord, F.M. The 'ability' scale in item characteristic curve theory. *Psychometrika* **1975**, *40*, 205–217. [CrossRef]
34. Bolt, D.M.; Deng, S.; Lee, S. IRT model misspecification and measurement of growth in vertical scaling. *J. Educ. Meas.* **2014**, *51*, 141–162. [CrossRef]
35. Feuerstahler, L.M. Sources of error in IRT trait estimation. *Appl. Psychol. Meas.* **2018**, *42*, 359–375. [CrossRef]
36. Scott, E.M.; Nolan, A.M.; Fitzpatrick, J.L. Conceptual and methodological issues related to welfare assessment: A framework for mmeasurement. *Acta Agric. Scand. Sect.—Anim. Sci.* **2001**, *51*, 5–10. [CrossRef]
37. Venham, L.L.; Gaulin-Kremer, E.; Munster, E.; Bengston-Audia, D.; Cohan, J. Interval rating scales for children's dental anxiety and uncooperative behavior. *Pediatr. Dent.* **1980**, *2*, 195–202. [PubMed]
38. Green, B. The method of successive intervals. In *Scaling*; Routledge: London, UK, 1974; pp. 122–128.
39. Torgerson, W.S. *Theory and Methods of Scaling*; Wiley: New York, NY, USA, 1958.
40. Luce, R.D.; Tukey, J.W. Simultaneous conjoint measurement: A new type of fundamental measurement. *J. Math. Psychol.* **1964**, *1*, 1–27. [CrossRef]
41. Avery, L.M.; Russell, D.J.; Raina, P.S.; Walter, S.D.; Rosenbaum, P.L. Rasch analysis of the gross motor function measure: Validating the assumptions of the Rasch model to create an interval-level measure. *Arch. Phys. Med. Rehabil.* **2003**, *84*, 697–705. [CrossRef]
42. Tennant, A.; Geddes, J.M.; Chamberlain, M.A. The Barthel index: An ordinal score or interval level measure? *Clin. Rehabil.* **1996**, *10*, 301–308. [CrossRef]
43. Borsboom, D.; Scholten, A.Z. The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory Psychol.* **2008**, *18*, 111–117. [CrossRef]
44. Kyngdon, A. The Rasch model from the perspective of the representational theory of measurement. *Theory Psychol.* **2008**, *18*, 89–109. [CrossRef]
45. Domingue, B. Evaluating the equal-interval hypothesis with test score scales. *Psychometrika* **2014**, *79*, 1–19. [CrossRef]

46. Karabatsos, G. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement Theory. *J. Appl. Meas.* **2001**, *2*, 389–423.
47. Shim, H.; Bonifay, W.; Wiedermann, W. Parsimonious Asymmetric item response theory modeling with the complementary log-log link. *Behav. Res. Methods* **2023**, *55*, 200–219. [CrossRef] [PubMed]
48. Mislevy, R.J. Chapter 6: Recent developments in item response theory with implications for teacher certification. *Rev. Res. Educ.* **1987**, *14*, 239–275. [CrossRef]
49. Krantz, D.; Luce, D.; Suppes, P.; Tversky, A. *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*; New York Academic Press: New York, NY, USA, 1971.
50. Ballou, D. Test scaling and value-added measurement. *Educ. Finance Policy* **2009**, *4*, 351–383. [CrossRef]
51. Gaito, J. Measurement scales and statistics: Resurgence of an old misconception. *Psychol. Bull.* **1980**, *87*, 564–567. [CrossRef]
52. Jamieson, S. Likert Scales: How to (ab) use them? *Med. Educ.* **2004**, *38*, 1217–1218. [CrossRef] [PubMed]
53. Lord, F.M. On the statistical treatment of football numbers. *Am. Psychol.* **1953**, *8*, 750–751. [CrossRef]
54. Marcus-Roberts, H.M.; Roberts, F.S. Meaningless statistics. *J. Educ. Stat.* **1987**, *12*, 383–394. [CrossRef]
55. Norman, G. Likert Scales, Levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ.* **2010**, *15*, 625–632. [CrossRef]
56. Prytulak, L.S. Critique of SS Stevens' theory of measurement scale classification. *Percept. Mot. Skills* **1975**, *41*, 3–28. [CrossRef]
57. Velleman, P.F.; Wilkinson, L. Nominal, ordinal, interval, and ratio typologies are misleading. *Am. Stat.* **1993**, *47*, 65–72. [CrossRef]
58. Robitzsch, A. On the bias in confirmatory factor analysis when treating discrete variables as ordinal instead of continuous. *Axioms* **2022**, *11*, 162. [CrossRef]
59. Boos, D.; Chen, J. Analysis of Likert-type data using metric methods. *Res. ONE* **2022**. *preprint*. Available online: https://researchers.one/articles/22.01.00002v1 (accessed on 31 March 2023).
60. Knapp, T.R. Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nurs. Res.* **1990**, *39*, 121–123. [CrossRef] [PubMed]
61. Fife-Shaw, C. Research methods in psychology: Approaches and methods. In *Research Methods in Psychology*; Breakwell, G.M., Wright, D.B., Smith, J.A., Eds.; Sage: Thousand Oaks, CA, USA, 2012; pp. 1–616.
62. Scholten, A.Z.; Borsboom, D. A reanalysis of Lord's statistical treatment of football numbers. *J. Math. Psychol.* **2009**, *53*, 69–75. [CrossRef]
63. Tal, E. Old and new problems in philosophy of measurement. *Philos. Compass* **2013**, *8*, 1159–1173. [CrossRef]
64. Tal, E. The Epistemology of Measurement: A Model-Based Account. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2012.
65. Van Fraassen, B.C. *Scientific Representation: Paradoxes of Perspective*; Oxford University Press: Oxford, UK, 2008.
66. Bringmann, L.F.; Eronen, M.I. Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory Psychol.* **2016**, *26*, 27–43. [CrossRef]
67. Chang, H. *Inventing Temperature: Measurement and Scientific Progress*; Oxford University Press: Oxford, UK, 2004.
68. Sherry, D. Thermoscopes, thermometers, and the foundations of measurement. *Stud. Hist. Philos. Sci. Part A* **2011**, *42*, 509–524. [CrossRef]
69. Consultative Committee for Thermometry. Mise En Pratique for the Definition of the Kelvin in the SI. Available online: https://www.bipm.org/documents/20126/41489682/SI-App2-kelvin.pdf/cd36cb68-3f00-05fd-339e-452df0b6215e (accessed on 3 April 2023).
70. Michell, J. Is psychometrics pathological science? *Measurement* **2008**, *6*, 7–24. [CrossRef]
71. McHugh, R.K.; Rasmussen, J.L.; Otto, M.W. Comprehension of self-report evidence-based measures of anxiety. *Depress. Anxiety* **2011**, *28*, 607–614. [CrossRef]
72. Santor, D.A.; Gregus, M.; Welch, A. FOCUS ARTICLE: Eight decades of measurement in depression. *Meas. Interdiscip. Res. Perspect.* **2006**, *4*, 135–155. [CrossRef]
73. Barrett, P. Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *J. Manag. Psychol.* **2003**, *18*, 421–439. [CrossRef]
74. Krabbe, P.F.; Stalmeier, P.F.; Lamers, L.M.; Busschbach, J.J. Testing the interval-level measurement property of multi-item visual analogue scales. *Qual. Life Res.* **2006**, *15*, 1651–1661. [CrossRef]
75. Blanton, H.; Jaccard, J. Arbitrary metrics in psychology. *Am. Psychol.* **2006**, *61*, 27. [CrossRef]
76. Ramsay, J.O.; Wiberg, M. A strategy for replacing sum scoring. *J. Educ. Behav. Stat.* **2017**, *42*, 282–307. [CrossRef]
77. Lee, M.K.; Peipert, J.D.; Cella, D.; Yost, K.J.; Eton, D.T.; Novotny, P.J.; Sloan, J.A.; Dueck, A.C. Identifying meaningful change on PROMIS short forms in cancer patients: A comparison of item response theory and classic test theory frameworks. *Qual. Life Res.* **2022**, 1–13. [CrossRef] [PubMed]
78. Johnston, B.C.; Ebrahim, S.; Carrasco-Labra, A.; Furukawa, T.A.; Patrick, D.L.; Crawford, M.W.; Hemmelgarn, B.R.; Schunemann, H.J.; Guyatt, G.H.; Nesrallah, G. Minimally important difference estimates and methods: A protocol. *BMJ Open* **2015**, *5*, e007953. [CrossRef]
79. Copay, A.G.; Subach, B.R.; Glassman, S.D.; Polly, D.W., Jr.; Schuler, T.C. Understanding the minimum clinically important difference: A review of concepts and methods. *Spine J.* **2007**, *7*, 541–546. [CrossRef]
80. Kolen, M.J.; Brennan, R.L. Score ccales. In *Test Equating, Scaling, and Linking*; Springer: Berlin, Germany, 2014; pp. 371–485.

81. Feuerstahler, L.M. Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika* **2019**, *84*, 105–123. [CrossRef] [PubMed]
82. Uher, J. Quantitative psychology under scrutiny: Measurement requires not result-dependent but traceable data generation. *Personal. Individ. Differ.* **2021**, *170*, 110205. [CrossRef]
83. Uher, J. Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences. *Qual. Quant.* **2022**, *56*, 2519–2548. [CrossRef]