Response to Moffat’s Comment on “Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales”

Marco Ferrante*, Nicola Ferro†, Norbert Fuhr‡

December 23, 2022

Abstract

Moffat recently commented on our previous work published in the IEEE Access journal. Our work focused on how laying the foundations of our evaluation methodology into the representational theory of measurement can improve our knowledge and understanding of the evaluation measures we daily use in Information Retrieval (IR) and how it can shed light on the different types of scales adopted by our evaluation measures; we also provided evidence, through extensive experimentation, on the impact and effect of the different types of scales on the subsequent statistical analyses, as well as on the impact of departing from their assumptions. Moreover, we investigated, for the first time in IR, the concept of meaningfulness, intended as a specific sort of invariance of the experimental statements and inferences you draw, and proposed it as a way to ensure more valid and generalizable results.

Moffat’s comments build on: (i) misconceptions about the representational theory of measurement, such as what an interval scale actually is and what axioms it has to comply with; (ii) they totally miss the central concept of meaningfulness. Therefore, we reply to Moffat’s comments by properly framing them in the representational theory of measurement and in the concept of meaningfulness.

All in all, we can only reiterate what we said several times: the goal of this research line is to theoretically ground our evaluation methodology – and IR is a field where it is extremely challenging to perform any theoretical advances – in order to aim for more robust and generalizable inferences – something we currently lack in the field. Possibly there are other and better ways to achieve this objective and these proposals could emerge from an open discussion in the field and from the work of others.

*Department of Mathematics “Tullio Levi-Civita”, University of Padua, Italy (email: ferrante@math.unipd.it)
†Department of Information Engineering, University of Padua, Italy (email: ferro@dei.unipd.it)
‡Faculty of Engineering, University of Duisburg-Essen, Germany (email: norbert_fuhr@unidue.de)
On the other hand, reducing everything to a contrast on what is (or pretend to be) an interval scale or whether all or none evaluation measures are interval scales may be more a barrier from than a help in progressing towards this goal.

1 Introduction

We write this response in reply to Moffat’s comment on our paper “Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales”.

We hope that this response will help to clarify some misunderstandings about the representational theory of measurement, in general, and the notion of meaningfulness in particular – misconceptions central to Moffat’s comment and leading to somehow fallacious arguments. We also hope that this response will contribute to a constructive and informed discussion in the field about these themes, which are often faced just as a contrast of opinions.

Information Retrieval (IR) is an highly experimental field by nature and by necessity, where we suffer from a lack of generalizability of our results and, consequently, the (almost) impossibility of predicting the performance of our systems. There is already some agreement on the need for “a better understanding of the assumptions and user perceptions underlying different metrics, as a basis for judging about the differences between methods” and on the fact that “the assumptions underlying our algorithms, evaluation methods, datasets, tasks, and measures should be identified and explicitly formulated. Furthermore, we need strategies for determining how much we are departing from these assumptions in new cases and how much this impacts on system performance”.

In this respect, our work adheres to the representational theory of measurement and proceeds step-by-step along its lines, in order to understand and check assumptions and deviations as well as their implications and consequences.

We welcome very much the comment by Moffat, as an exemplar way: (i) to open the discussion to the community; (ii) to keep a written record of the arguments, record which, today, helps researchers in taking informed decisions and, in the future, will remain useful to motivate adopted approaches or to further revise them; and, (iii) to have a transparent and frank exchanges of views on a somewhat heartfelt topic. On the other hand, we would very much prefer all of this not to be framed as a contrast between, in Moffat’s words, “a bleak picture of past decades of IR evaluation” and “a more optimistic view of IR evaluation and IR measurement”, because this rhetoric is apt to fuel oppositions, leaving everything unchanged, rather than to tackle the real issue, perhaps to favor the formulation of alternative and better solutions, and to make overall progress.

The paper is organized as follows: Section summarizes the main arguments in Moffat’s comment; Section introduces preliminary notions on the representation theory of measurement and on meaningfulness in order to set a proper stage for the discussion; Section provides our response to Moffat’s
main arguments; finally, Section 5 reports some concluding remarks.

2 Main Arguments by Moffat

As Moffat’s comment does not refer to any specific part of our paper to which we could answer, we will start from his text instead, considering his main arguments by directly quoting them. We also limit ourselves to the very major arguments, avoiding to enter in too low level or too technical details, because this would make our response too long, less clear, and of little use to a researcher who wants to form their opinion or to contribute new ideas.

**MA1** “If you do know where the numbers came from and why they have the values that they do, and are confident that those values can be justified in reference to the real world attribute that the mapping is designed to represent, then those numbers may be used in your analysis and interpretation of that real world equivalence” [20, p. 105570]

**MA1.1** “While care needs to be exercised when choosing the metric that best fits the user experience for any particular IR application [...] once that match has been decided, the values calculated by the effectiveness metric may be used as simple numbers “that don’t remember where they came from” [18]; that is, without regard to their origins in a categorical-scale SERP dataset” [20, p. 105576]

**MA2** “Via a sequence of examples we have presented our view that all IR effectiveness metrics can be considered to be interval scale measurements, provided only that the mapping from SERP categories to numeric scores has a real-world basis (an external validity) and can be motivated as corresponding to the underlying usefulness of each SERP, as experienced by an identified cohort of users as they carry out some identified search task” [20, p. 105576]

**MA2.1** “There can be no ambiguity: RR is an interval scale measurement for those users” [20, p. 105573]

**MA2.2** “Any argument that RR – or any other metric – is an unsuitable categorical to numeric mapping for measuring IR system effectiveness for some cohort of users or some type of search task must be justified based on rhetoric about user perceptions of SERP usefulness, or on observational data that measures SERP usefulness via some agreed surrogate. Arguments against IR effectiveness metrics cannot be based solely upon statements about the non-uniformity of the intervals between the available measurement points” [20, pp. 105573–105574]

**MA2.3** “We argue that most current IR metrics are well-founded, and, moreover, that those metrics are more meaningful in their current form than in the proposed ‘intervalized’ versions” [20, p. 105564]
MA3 “We believe that intervalization should regarded with scepticism. There is no requirement in Steven’s typology that interval scales be restricted to uniform distances between the available measurement points; the requirement is simply that the ratio between pairs of intervals be indicative of the corresponding difference in the underlying attribute” [20, p. 105574]

MA3.1 “We have argued that the proposed intervalization of current IR effectiveness metrics is neither required nor helpful. If the raw metric value is indeed a defensible measurement of SERP usefulness and corresponds to the user’s experience when they are presented with a member of that SERP category, then equi-intervalizing those measurements via a different categorical to numeric mapping must of necessity distort and alter any findings that arise, and thus risk masking what would otherwise be valid conclusions. And if the raw metric is not a defensible measurement of SERP usefulness for the search task at hand, then equi-intervalizing its scores is unlikely to improve the situation” [20, p. 105576]

MA3.2 “Moreover, altering the categorical to numeric mapping used to assign score to SERPs changes the relativities being measured, and thus affects the outcome of any subsequent arithmetic” [20, p. 105574]

3 Preliminaries on Measurement and Meaningfulness

Although we spent quite some effort to explain the following concepts in our original article [12], as well as in previous and more formal work [10], providing detailed discussion, references, and examples, we summarize here the main concepts needed to articulate our response. In doing so, we also opt for quoting as many passages as possible directly from the foundational works in the area of measurement, in order to plainly report third party research, without any additional interpretation on our side.

An introductory presentation of the concepts described below can be found in general textbooks about measurement, such as Hand [15], or textbooks about software metrics, such as Fenton and Bieman [7]. For a historical overview of the evolution of the theory of measurement and its concepts, you can refer to Díez [5, 6].

3.1 Measurement and Scales

“When measuring some attribute of a class of objects or events, we associate numbers [...] with objects in such a way that the properties of the attribute are faithfully represented as numerical properties” [17, p. 1].

Suppose we aim at measuring the length of rods. In the real world, we can compare rods to determine which one is longer, i.e. we have a comparison
relation \( \succ \) among rods; we can also concatenate rods together, i.e. we have a concatenation operation \( \circ \) among rods.

“A relational structure is a set with one or more relations on that set” [17, p. 8].

So, in our example, we have a set of rods \( A \), a binary relation \( a \succ b \) to compare them, and a ternary relation \( c = a \circ b \) to concatenate them. Overall, \( \langle A, \succ, \circ \rangle \) is an empirical relational structure over the set of rods; it is called empirical because it exists in the real world.

“The numerical assignment \( \phi \) is a homomorphism in the sense that it sends \( A \) into \( \mathbb{R}^+ \), \( \succ \) into \( > \), and \( \circ \) into \( + \) in such a way that \( > \) preserves the properties of \( \succ \) and \( + \) the properties of \( \circ \)” [17, p. 8], where the properties are [17, p. 4]:

1. \( a \succ b \) if and only if \( \phi(a) > \phi(b) \);
2. \( \phi(a \circ b) = \phi(a) + \phi(b) \).

These small excerpts from Krantz et al. [17], the first volume of the landmark three-volume series on the foundations and mathematical formalization of measurement, provide us with an intuitive understanding of the basic idea behind the measurement theory: we use numbers as proxies of attributes of real world objects, provided that the relations and operations among those numbers keep corresponding to the relations and operations among real world objects.

This is formulated mathematically by saying that the assignment \( \phi \), which is called a scale, is, in general, a homomorphism. Note that the homomorphism is used because \( \phi(a) = \phi(b) \) does not necessarily mean that the rods \( a \) and \( b \) are the same rod but just that they have the same length. However, if we consider the equivalence relation \( \sim \) on \( A \), then the empirical relational structure \( \langle A/\sim, \succ, \circ \rangle \) on the quotient set \( A/\sim \) becomes an isomorphism to the numerical relational structure \( \langle \mathbb{R}^+, \geq, + \rangle \). This further underlines the bijective nature of the correspondence between real world objects and numbers, as well as their relations and operations.

The theory of measurement moves a further step forward and asks and additional question: “Given a set of rods, a comparison relation \( \succ \), and a concatenation operation \( \circ \), what assumptions concerning \( \succ \) and \( \circ \) are necessary and/or sufficient to construct a real-valued function \( \phi \) that is order preserving and additive?” [17, p. 8]. Therefore, we have to seek for characteristics of the real world objects, expressed in the form of axioms, which guarantee the existence of a numerical scale \( \phi \) with the desired properties. More specifically, “a representation theorem asserts that if a given relational structure satisfies certain axioms, then a homomorphism into a certain numerical relational structure can be constructed [...] Measurement can be regarded as the construction of homomorphisms (scales) from empirical relational structures of interest into numerical relational structures that are useful” [17, p. 9].

However, the procedure for numerical assignment may look somewhat arbitrary or there may exist several alternatives which are equally good. For example, in the case of rods, which rod size is chosen as unit is an arbitrary matter, leading to equivalent scales. In other cases, it may be less immediate to
see which choices are arbitrary and which are not. For example, when we count how many times the unit rod $u$ has to be concatenated for obtaining a given rod $a$, why do we record $n$ instead of $n^2$ or $e^n$?

The notion of permissible transformation serves exactly the purpose of answering this question: “A transformation $\phi \rightarrow \phi'$ is permissible if and only if $\phi$ and $\phi'$ are both homomorphisms of $\langle A, R_1, \ldots, R_n \rangle$ into the same numerical structure $\langle \mathbb{R}, S_1, \ldots, S_n \rangle$” [17, p. 12]. An uniqueness theorem has the purpose of determining what this permissible transformation is and this is often not obvious at all.

For example, in the case of rod and length, Hölder [16] developed the first proof that if $\phi$ is order preserving and additive, i.e. a homomorphism of $\langle A, \succ, \circ \rangle$ into $\langle \mathbb{R}^+, \succ, + \rangle$, the same is true for $\alpha \phi$ when $\alpha > 0$; moreover, if $\phi'$ is any homomorphism of $\langle A, \succ, \circ \rangle$ into $\langle \mathbb{R}^+, \succ, + \rangle$, then $\phi' = \alpha \phi$ for some $\alpha > 0$.

The uniqueness theorem grasps the fact that even if we refer to different scales, e.g. meters or feet for length, they are actually equivalent from the standpoint of the permissible transformation and they are required to be so, otherwise the homomorphism with the real world $\langle A, \succ, \circ \rangle$ would be lost.

In this context, Stevens [25] defined the different types of scales, i.e. different types of $\phi$ which have to comply with specific axioms in order to guarantee desired properties. They can be briefly summarized as follows:

- **Nominal scale**: it is used when entities of the real world can be placed into different classes or categories on the basis of their attribute under examination, without any notion of ordering among them. Any distinct numeric representation of the classes is an acceptable measure but there is no notion of magnitude associated with numbers.

Therefore, any arithmetic operation on the numeric representation has no meaning. As a consequence, the only allowable statistics is counting number of items in each class, that is mode and frequency.

The class of permissible transformations is the set of all one-to-one mappings, i.e. bijective functions: $\phi' = f(\phi)$, since they preserve the distinction among classes.

- **Ordinal scale**: it can be considered as a nominal scale where, in addition, there is a notion of ordering among the different classes or categories. Any distinct numeric representation which preserves the ordering is acceptable. Therefore, the magnitude of the numbers is used just to represent the ranking among classes.

Addition, subtraction or other mathematical operations have no meaning. As a consequence, besides the statistics already allowed for nominal scales, median, quantiles, and percentiles are appropriate, since there is a notion of ordering.

The class of permissible transformations is the set of all the monotonic increasing functions, since they preserve the ordering: $\phi' = f(\phi)$. 


• **Interval scale**: besides relying on ordered classes, it also captures information about the size of the intervals that separate the classes. An interval scale preserves order, as an ordinal one, and differences among classes have meaning – but not their ratio.

Addition and subtraction are acceptable operations but not multiplication and division. As a consequence, besides the statistics allowed for ordinal scales, mean and standard deviation are allowable since they depend just on sum and subtraction.\(^1\)

The class of permissible transformations is the set of all the *affine transformations*:\( \phi' = \alpha \phi + \beta, \alpha > 0.\)

Temperature is the typical example of an interval scale.

• **Ratio scale** it allows us to compute ratios among the different classes since classes are ordered.

All the arithmetic operations are allowed. As a consequence, besides the statistics allowed for interval scales, geometric and harmonic mean, as well as coefficient of variation, are allowable since they depend on multiplication and division.

The class of permissible transformations is the set of all the *linear transformations*:\( \phi' = \alpha \phi, \alpha > 0.\)

Length and mass are typical examples of ratio scales.

Let us focus on interval scales, which are the matter of discussion in our paper and in Moffat’s comment.

The important characteristics of interval scales is that they need to be based on *equally spaced* objects in the real world which, in turn, leads to equi-spacing of the numerical mapping.

Stevens does not explicitly mention the term “equi-spacing” in Table 1 at page 678, where he summarizes the scale types. However, he explicitly says this in the section where he explains what interval scales are: “most 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”\(^2\), providing also concrete examples like “equal intervals of temperature are scaled off by noting equal volumes of expansion”\(^2\). Other sentences like “the scale form remains invariant when a constant is added”\(^2\) or “if the purpose of the scale is still served when its values are squared or cubed, it is not even an interval scale”\(^2\) implicitly assume an equally spaced scale.

Let us now see how an interval scale is defined from a formal point of view. Rossi\(^3\) explains “How can we assign numbers (measures) to them [objects] in such a way that they properly express both the order and the distances?\(^4\)

\(^1\)Note that when we talk about admissible operations, we mean operations between items of the scale. So, for example, a mean involves summing items of the scale, e.g. temperature, and this is possible on an interval scale. The fact that a mean also requires a division by the number \(N\) of items added together is not in contrast with saying that only addition and subtraction are allowed, since \(N\) is not an item of the scale.
For doing so, we have to establish an *equally spaced graduation*. In particular, the equally spaced graduation is formulated in terms of a so-called *solvability condition*, requiring that whenever we have two not equivalent intervals, it is always possible to find elements which correspond to each treat of the equally spaced graduation Rossi [22, p. 57].

All the properties required required to real world object to allow for the creation of an interval scale are determined by the definition of a *difference structure*, i.e. specific type of empirical relational structure. Axiom 12.4 of Definition 3.12 (Difference Structure) [22, p. 59] expresses the *solvability condition*:

\[
\text{if } \Delta_{ab} \succeq_d \Delta_{cd} \succeq_d \Delta_{aa} \text{ then there exist } d' \in A \text{ and } d'' \in A \\
\text{so that } \Delta_{ad'} \sim_d \Delta_{cd} \sim_d \Delta_{d'b}
\]

where \( \Delta_{ab} \) is the *difference* in the real world, not among numbers, and \( \succeq_d \) is a weak order among differences, again in the real world.

The definition of difference structure is then used in the *Representation Theorem* 3.17 to demonstrate the existence of an interval scale [22, p. 62]:

\[
\Delta_{ab} \succeq_d \Delta_{cd} \iff \phi(a) - \phi(b) \geq \phi(c) - \phi(d)
\]

based on the equivalence \( \iff \) between the notion of difference in the real world, consisting of equi-space objects, and the notion of different in the numerical system.

Finally, the *Uniqueness Theorem* 3.18 demonstrates that the affine transformation \( \phi' = \alpha \phi + \beta, \alpha > 0 \), is the permissible transformation for an interval scale [22, p. 62].

A similar formalization is adopted even more extensively by Krantz et al. [17, pp. 136ff.].

As a consequence of the above formal definition of interval scale, the ratio of differences among classes, i.e. the ratio of intervals, is allowed and invariant to an affine transformation [17, p. 10]:

\[
\frac{\phi'(a) - \phi'(b)}{\phi'(c) - \phi'(d)} = \frac{[\alpha \phi(a) + \beta] - [\alpha \phi(b) + \beta]}{[\alpha \phi(c) + \beta] - [\alpha \phi(d) + \beta]} = \frac{\phi(a) - \phi(b)}{\phi(c) - \phi(d)}
\]

### 3.2 Meaningfulness

*Meaningfulness* is a technical term in the theory of measurement, which relies on a precise mathematical formulation and serves the purpose of developing a whole theory around it. It should not be confused with “meaningful” in the everyday language sense, i.e. making sense or being credible.

For a rigorous and formal definition of *meaningfulness*, please, refer to Narens [21]. For the purpose of the present discussion, the intuitive definition by Adams et al. [1, pp. 99-100] should suffice: “the criterion of appropriateness for a statement about a statistical operation is that the statement be empirically meaningful in the sense that its truth or falsity must be invariant under permissible transformations of the underlying scale".
Therefore, meaningfulness focuses on the invariance of the statements we make and not on how much sense the make to us or how much true or false they are. A statement like “A Chihuahua dog is three times taller than a Great Dane dog” is false (possibly it does not make much sense either) and it stays false independently from whether we are using meters or feet, i.e. under a permissible linear transformation of the scale.

Meaningfulness is not different from the notion of invariance we have in geometry, when we ask that a shape remains the same independently from translation or rotation. We may wonder if meaningfulness is asking too much or a too strong property to a scale. But, asking that the truth value of a statement remains the same under permissible transformations, which we have seen to be an intrinsic and indispensable property of a scale, is the same as asking that we draw the same inferences and conclusions independently from whether we are using meters or feet. All of this does not seem much stricter than asking, when we are looking at a cube, that, if we rotate it by 30 degrees, we still see a cube and not a sphere instead.

To clarify how meaningfulness works, for example in the case of an interval scale, we report here an example taken from our previous paper [12].

**Example 1 (Meaningfulness for an Interval Scale).** The statement “Today the difference in temperature between Rome and Oslo is twice as high as it was one month ago” is meaningful. Indeed, if, on the Celsius scale, the temperature today in Rome is $20\,^\circ C$ and in Oslo is $10\,^\circ C$ while one month ago it was $12\,^\circ C$ and $7\,^\circ C$, leading to $20 - 10 = 10$ which is twice as $12 - 7 = 5$, on the Fahrenheit scale we would have $68 - 50 = 18$ which is twice as $53.6 - 44.6 = 9$.

Suppose now that we have recorded two sets of temperatures from Paris and Rome: $T_C^P = [2, 4, 8, 36]$ and $T_C^R = [1, 2, 4, 15, 34]$ in Celsius degrees and, the same, $T_F^P = [35.6, 35.6, 39.2, 46.4, 96.8]$ and $T_F^R = [33.8, 35.6, 39.2, 59.0, 93.2]$ in Fahrenheit degrees.

The statement “The median temperature in Paris is the same as in Rome” is meaningful, since $4 = 4$ in Celsius degrees and $39.2 = 39.2$ in Fahrenheit degrees; this is due to the fact that interval scales are also ordinal scales and quantiles are an allowable operation on ordinal scales.

The statement “The mean temperature in Paris is less than in Rome” is meaningful as well, since $10.4 < 11.2$ in Celsius degrees and $50.72 < 52.16$ in Fahrenheit degrees; this is due to the fact that addition and subtraction are allowable operations on an interval scale and, as a consequence, mean is allowable as well.

Finally, the statement “The geometric mean of temperature in Paris is greater than in Rome” is not meaningful, since $5.40 > 5.27$ in Celsius degrees and $46.74 < 48.17$ in Fahrenheit degrees; this is due to the fact that the geometric mean involves the multiplication and division of values, which is not a permitted operation on an interval scale.

Examples along this lines could be done for an ordinal scale showing that the median (or any other percentile) remains meaningful for any monotonic increasing transformation but not the mean (or even the geometric mean).
4 Response to Main Arguments by Moffat

4.1 MA1: If you do know where the numbers came from . . .

We all agree that numbers should be chosen in such a way to reflect the attribute of a real world object and, in this respect, taking into account also the user experience does not make an exception.

However, at the same time, from the discussion in Section 6.1 it should be clear that operations among the numbers should keep corresponding to operations among real world objects and that numbers should keep their link with the attribute of the real world object.

More formally, the theory of measurement explicitly says that you have to create an homomorphism between the real world and the numerical system in order to ensure that relations and operations are preserved. Even more, the theory of measurement makes a step further and states that among real world objects certain properties should hold in order to ensure both the existence (representation theorem) of a mapping to numbers with desired properties and its uniqueness (uniqueness theorem). This is, for example, the case of the difference structure [22, Section 3.4, pp. 55ff.] and [17, Chapter 4, pp. 136ff.] which impose axioms among the real world objects that allow for constructing an interval scale.

Moffat reported the famous statement by Lord [18] who, via the opinion of a statistician in his story, says: “since the numbers don’t remember where they came from, they always behave just the same way, regardless”. This statement is often used to support the use of whatever operation on numbers, regardless of any scale consideration and justified by practical usefulness. It is well known that Lord’s paper has been debated for decades and we reported the different viewpoints on it in our original paper [12].

In particular, Scholten and Borsboom [24], referred also by Moffat, show that Lord’s argument is broken and you cannot compute a mean on an ordinal scale just because “numbers don’t remember where they come from”. On the contrary, Scholten and Borsboom demonstrate, also formally in their appendix, that the reason why the means computed by the statistician in Lord’s paper can work is completely different and not related to Lord’s statement. Indeed, the question these means are answering is not about “football players numbers”, as Lord assumes, but about the bias in the machines assigning those numbers to players and how they have been tampered with. In other terms, the means are about a different attribute of a different real world object (the machines and not the football players) and Scholten and Borsboom showed that on this new attribute it is possible to define a bisymmetrical structure which, in turn, leads to an interval scale. Therefore, Scholten and Borsboom demonstrate that even the Lord’s argument, when properly conceptualized, leads to and supports Stevens’s definition of scales and the admissible operations on them.

Note that the bisymmetrical structure used by Scholten and Borsboom to define an interval scale relies on a solvability condition (equi-spacing) like those used in the difference structures mentioned in Section 6.1.
Moreover, Moffat improperly uses the term *meaningfulness* in its everyday language sense instead of its scientific one and this leads to wrong conclusions. Sentences like “that factual relationship makes the mapping’s values *meaningful*, and hence interpretable in terms of the attribute from which the measurement was derived” [20, p. 105569] or “the intervals between the salary points are *meaningful*; they represent salary differentials that must be paid in a competitive market, measured in dollars” [20, p. 105569] are wrong and, certainly, do not support Moffat’s thesis about the scales being created. Indeed, as explained, *meaningfulness* is a form of invariance of the statements you make; it is not some “quality” of the numbers you assign and, certainly, it does not express how much sense those numbers may have for you.

Overall, we think that MA1 is not correct and does not hold since, for all the reasons explained above: even “if you do know where the numbers came from”, this is not enough to them manipulate them as if “numbers don’t remember where they come from”. In other terms, you should still take into consideration the properties of the scale those numbers belong to and admissible operations on that scale.

4.2 MA2: All IR effectiveness metrics can be considered to be interval scale measurements

This argument by Moffat revolves around two severe misunderstandings.

The first misconception is that an interval scale does not require equally spaced steps and, thus, whatever IR measure, even not equally spaced, can pretend to be an interval scale. In Section 3.1, we have explained how the *solvability condition*, which express equi-spacing, is one of the axioms required to define an interval scale. Therefore, whatever numerical mapping not complying with the solvability condition does not match the definition and cannot be called an interval scale. On the other hand, a scale where order is preserved but intervals are not equi-spaced exists and it is an *ordinal scale*, as also remarked by Stevens [25, p. 679]: “on an ordinal scale [...] the successive intervals on the scale are unequal in size”.

The second misconception is about what *meaningfulness* is in the theory of measurement. Moffat just uses *meaningfulness* in its everyday language use of ‘making sense’ or ‘being credible’ along examples like when, in Sections II-D and II-E of his paper, the provost analyses the professor salaries or like when, in Section III-D of his paper, a market study leads to associate a Search Engine Result Page (SERP) with a numerical mapping corresponding to Reciprocal Rank (RR). All of this has nothing to do with the notion of *meaningfulness* in the theory of measurement and, as it should be clear from the discussion in the previous section, having some rationale in assigning numbers to objects does not imply or guarantee any scale properties, since much more precise conditions should be met in this case, such as the solvability condition. Therefore, a sentence like “most current IR metrics are well-founded, and, moreover, [...] those metrics are more *meaningful* in their current form” [20, p. 105564] is wrong because *meaningfulness* neither is a property of a measure nor it is a synonym
of well-founded nor it something you can have “more” or “less”, because either a statement is meaningful (invariant) or it is not. Even more, it is not a transitive property you can use to justify some sort of chain of reasoning like: the assignment of numbers makes sense to me (this is not meaningful), thus the measure makes sense to me (this is not meaningful), thus the measure is an interval scale (this is neither meaningful nor being an interval scale), thus the operations with that numbers make sense (this is not meaningful), thus the statements/inferences make sense (here, in case, meaningful would be about the invariance of the statement and not their sense or truth).

Overall, we think that MA2 is not correct and does not hold since, for all the reasons explained above: neither all IR measures are interval scales nor can you always compute means over them and obtain meaningful statements in a scientific sense nor interval scales with intervals not equi-spaced exist.

As a side note, for these reasons, the examples by Moffat on professor salaries and SERP pages/RR are not interval scales, as a consequence means cannot be computed and the resulting statement cannot be meaningful.

4.3 MA3: We believe that intervalization should be regarded with scepticism

The intervalization procedure we proposed in our paper proceeds along this lines: (i) generate all the possible SERP; (ii) compute the desired IR evaluation measure; (iii) sort the SERP according to the computed measure; (iv) use the rank assigned to a SERP in (iii), which is an interval scale by construction, in the subsequent analyses and statistical test. Since, as already discussed in Section 4.2, not all the IR evaluation measures can be considered to be interval scales, the proposed intervalization could find some useful application.

As we already discussed in Section 3.1 the solvability condition, i.e. equally spaced steps, is an axiom to be complied with for having an interval scale. Moreover, despite what Moffat reports about Stevens’s paper, Stevens actually says that equi-spacing is a requirement for an interval scale. Therefore, not holding, this should not be taken as a motivation for regarding the proposed intervalization with scepticism.

In our paper, we do not claim that the purpose of intervalization is to make an IR evaluation measure a more “defensible measurement of SERP usefulness”. We actually assume that every IR evaluation measure embeds its own user viewpoint – being it defensible or not – and, in case that a measure does not comply with the requirements for an interval scale, our intervalization procedure tries to preserve that user viewpoint as much as possible, still obtaining an interval scale, by keeping the same ordering of SERP produced by that user viewpoint. Therefore, not holding, this should not be taken as a motivation for regarding the proposed intervalization with scepticism.

The fact that the intervalization procedure changes the numerical mapping and that this will affect the subsequent computations is quite a trivial observation and it is exactly the purpose of this transformation. Indeed, our paper
reports an extensive experimentation and a throughout investigation of the im-

pacts and effects of this transformation on several kinds of statistical analyses

and across several standard test collections.

Overall, what our intervalization procedure gives you is the possibility of for-

mulating meaningful statements, doing its best at preserving the user view-

point embedded by an IR evaluation measure; it does not aim at all at making

that user viewpoint more or less defensible. Therefore, researchers may or may

not be interested in this approach and may or may not adopt it in their own

experiments; researchers may hopefully also come up with other more brilliant

solutions to ensure the meaningfulness of our statements. Perhaps, researchers

will not be sceptic about our intervalization procedure for the very same reasons

suggested by Moffat or, maybe, Moffat will provide some experimental evidence
to substantiate his scepticism and, in case, to help identifying and quantifying

specific issues and how to address them.

4.4 Other Considerations

Our line of work

There is some misunderstanding in Moffat’s comment about how different lines

of our work relate together, often mentioning them all together as if they were

all the same, reporting the same claims, or as if all of them could be refuted by

the same argument.

Fuhr [14] listed a series of practices he considers common mistakes in IR

evaluation, among which averaging RR since it is not an interval scale, to be

avoided. Sakai [23] already commented on the prescriptive nature of Fuhr’s

paper and on what he considers or not to be mistakes; among them, Sakai

considers averaging RR a legitimate operation.

Our work commented by Moffat actually originates from a different line of

research. In our first work [8], we started to seek for a way to apply the theory

of measurement to the IR evaluation by finding axioms (swap and replacement

among relevant documents in a SERP) that described the properties of a SERP

in the real word, a prerequisite for understanding the properties of IR evaluation

measures and their scales. Later on, in [9] and, more completely, in [10] we used

the notion of difference structure explained in Section 3.1 to formally describe

the SERP in the real world and from there to derive an interval scale measure;

this, in turn, allowed for verifying which IR evaluation measures were an inter-

val scale (at least with respect to the identified difference structure) by seeking

for an affine transformation. Then, in [11] we started to explore the impact

that using or not an interval scale can have in IR experimentation. Finally, in

our last work [12] commented by Moffat, we joined our interests and proposed

meaningfulness as a fundamental concept to be accounted for in IR evaluation,

we further investigated the implications of IR measures being interval scales or

not and why, and we proposed intervalization as a viable approach to ensure

meaningfulness. All in all, the objective of all these works is to provide bet-

ter theoretical foundations to our evaluation methods and concrete means to
achieve them. In this respect, these works are in the wake that others have also followed but using different approaches, such as van Rijsbergen [27], Bollmann [2], Bollmann and Cherniavsky [3], or Amigó and Mizzaro [2].

Finally, to the best of our knowledge, Ferrante et al. [11, 12] represent the first works to experimentally investigate in a systematic way the impact of scales and departure from their assumptions on different types of statistical analyses.

In this perspective, the opening statement in the abstract of Moffat’s paper “A sequence of recent papers, including in this journal, has considered the role of measurement scales in information retrieval (IR) experimentation, and presented the argument that (only) uniform-step interval scales should be used. Hence, it has been argued, well-known metrics such as reciprocal rank, expected reciprocal rank, normalized discounted cumulative gain, and average precision, should be either discarded as measurement tools, or adapted so that their metric values lie at uniformly-spaced points on the number line” is, at best, reductive of what our work actually is and shifts the focus of the discussion from laying theoretical and experimental foundations for our experimental methodology to “use/do not use interval scales” or ”use/do not use that evaluation measure”.

Recall Base

When talking about the problems caused by the recall base (RB), i.e. the number of relevant documents for a topic, in Section IV-B of his paper, Moffat states that “our contention in this work is that the measurement scale is always the positive real number line, and hence that no question of alignment (or not) of measurement points across sets of topics arises” since, always according to Moffat, “from the point of view of Fuhr and Ferrante et al., those difficulties arise because normalization by RB means that the set of generable measurement points for any query in a set of topics might not numerically align with the available measurement points for other topics that have different values for RB”.

Actually, the view point expressed in our paper is that measures which explicitly depend on the recall base in their formulation lead to different scales on different topics and, as a consequence, they cannot be compared or mixed up, like you would not mix up length and mass, even if they are both ratio scales. So, the issue is more profound than a lack of numerical alignment.

Finally, in Section III-G of his paper, Moffat suggests that, differently from what reported in our previous works, also Rank-Biased Precision (RBP) with \( p = 0.5 \) is not an interval scale when, for example, a SERP is truncated at a length \( k \) greater than the recall base for that topic. Actually, what our previous work shows is that RBP with \( p = 0.5 \) is an interval scale when you construct it by assuming that as many relevant documents as needed are available, independently from the length of the SERP or its truncation point. This is quite a reasonable assumption because you are creating a scale which should hold for all the topics and not a different RBP scale for topics with one relevant document, topics with two relevant documents, and so on. It is the same line of reasoning you adopt when creating a scale for length: neither you create a scale for a
geographical area with small trees and a separate one for another area with tall
trees, nor you attempt to remove from a scale the ticks corresponding to some
trees for discovering later on that then it is no more an interval scale.

As a side note, from a more formal point of view, as discussed in Section 3.1,
the solvability condition \[22, pp. 56ff.\] and \[17, pp. 136ff.\], which leads to equally
spaced steps, is one of the properties required for constructing an interval scale
and, if your real world systems do not match it, e.g. because they are limited to
the case of a single relevant document, it trivially follows that you can construct
an interval scale. However, as explained before, this is not an issue of the
numerical mapping but rather of the real world objects that lack the needed
properties to create an interval scale.

5 Concluding Remarks

In this paper we have replied to Moffat’s comment on our previous work \[12\]. In
doing so, we have briefly summarised the main concepts of the representational
theory of measurement \[17, 19, 26\] and of meaningfulness \[21\] and we have
explained why the main arguments by Moffat do not hold, mainly because of
misconceptions on these foundational concepts of the representational theory of
measurement and of meaningfulness.

As said, we really welcome Moffat’s comment on our work because it offers
the opportunity for an open discussion on important topics for our field. On the
other hand, we remark that we would prefer avoiding to frame the discussion as
the contrast between “a bleak picture of past decades of IR evaluation”
and “a more optimistic view of IR evaluation”, suggested by Moffat. First,
in our work we have never criticized or even deprecated past decades of IR
evaluation. Saying that scale properties of evaluation measures matter or that
meaningfulness matters is neither criticizing nor deprecating past research, at
least any more than talking about modern building techniques can be seen as a
criticism of the pyramids. Nor we can consider as a criticism asking our selves
which statements in the IR literature are meaningful, i.e. invariant, because
this, for example, would help in knowing what could potentially generalize in
an easier way. Second, framing the question as a contrast risks to amplify a
“defensive attitude” in the field, rightly motivated by safeguarding the seminal
and paramount results of our past research, at the expense of an open-minded
discussion of the topic and of a collaboration among researchers on how to
develop and adopt better foundations for our evaluation methods.

A general feeling that emerges from Moffat’s comment is that we overlook the
user viewpoint or we do not account for the user experience. On the contrary,
we very much agree with Moffat and all the rest of the research community
that the user satisfaction is the ultimate goal of our measurement and that
evaluation measures should embed some user viewpoint. We just say that we
should strive for this goal in the most sound and safe way possible; the proposed
intervalization procedure is just a simple example of such an attempt. We
actually do hope that this discussion and, especially, further research by others
will deliver much better solutions in this respect.

Both here and in our work commented by Moffat, we have indicated what this “sound and safe way” could be, i.e. a deeper investigation and adoption of the concept of meaningfulness, at least in its scientific sense of invariance. Indeed, meaningfulness could be a way the achieve that generalizability of results that we lack in our field. Moreover, meaningfulness frees us from the debate on “should we average or not?” or “should we give or adhere to prescriptions or not?” and focus our attention on the real goal, i.e. drawing more robust and generalizable inferences and conclusions through better foundations of our evaluation methods.

Finally, independently from the stance researchers may have on these topics, we think that more thorough experimentation should be carried out in the field also by others, in order to transfer theoretical models and considerations to practice, to gain a better understanding of the implications of the different choices, and to have a more informed discussion on the pros and cons of the various alternatives.

References

[1] E. W. Adams, R. F. Fagot, and R. E. Robinson. A theory of appropriate statistics. *Psychometrika*, 30:99–127, June 1965.

[2] E. Amigó and S. Mizzaro. On the nature of information access evaluation metrics: a unifying framework. *Information Retrieval Journal*, 23(3):318–386, June 2020.

[3] P. Bollmann. Two Axioms for Evaluation Measures in Information Retrieval. In C. J. van Rijsbergen, editor, *Proc. of the Third Joint BCS and ACM Symposium on Research and Development in Information Retrieval*, pages 233–245. Cambridge University Press, UK, 1984.

[4] P. Bollmann and V. S. Cherniavsky. Measurement-theoretical investigation of the MZ-metric. In C. J. van Rijsbergen, editor, *Proc. 3rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1980)*, pages 256–267. ACM Press, New York, USA, 1980.

[5] J. A. Díez. A Hundred Years of Numbers. An Historical Introduction to Measurement Theory 1887-1990. Part I: The Formation Period. Two Lines of Research: Axiomatics and Real Morphisms, Scales and Invariance. *Studies in History and Philosophy of Science*, 28(1):167–185, March 1997.

[6] J. A. Díez. A Hundred Years of Numbers. An Historical Introduction to Measurement Theory 1887-1990. Part II: Suppes and the Mature Theory. Representation and Uniqueness. *Studies in History and Philosophy of Science*, 28(2):237–265, June 1997.
[7] N. E. Fenton and J. Bieman. *Software Metrics: A Rigorous & Practical Approach*. Chapman and Hall/CRC, USA, 3rd edition, 2014.

[8] M. Ferrante, N. Ferro, and M. Maistro. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In J. Allan, W. B. Croft, A. P. de Vries, C. Zhai, N. Fuhr, and Y. Zhang, editors, Proc. 1st ACM SIGIR International Conference on the Theory of Information Retrieval (*ICTIR 2015*), pages 21–30. ACM Press, New York, USA, 2015.

[9] M. Ferrante, N. Ferro, and S. Pontarollo. Are IR Evaluation Measures on an Interval Scale? In J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, editors, Proc. 3rd ACM SIGIR International Conference on the Theory of Information Retrieval (*ICTIR 2017*), pages 67–74. ACM Press, New York, USA, 2017.

[10] M. Ferrante, N. Ferro, and S. Pontarollo. A General Theory of IR Evaluation Measures. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 31(3):409–422, March 2019.

[11] M. Ferrante, N. Ferro, and E. Losiouk. How do interval scales help us with better understanding IR evaluation measures? *Information Retrieval Journal*, 23(3):289–317, June 2020.

[12] M. Ferrante, N. Ferro, and N. Fuhr. Towards Meaningful Statements in IR Evaluation. Mapping Evaluation Measures to Interval Scales. *IEEE Access*, 9:136182–136216, 2021.

[13] N. Ferro, N. Fuhr, G. Grefenstette, J. A. Konstan, P. Castells, E. M. Daly, T. Declerck, M. D. Ekstrand, W. Geyer, J. Gonzalo, T. Kuflik, K. Lindén, B. Magnini, J.-Y. Nie, R. Perego, B. Shapira, I. Soboroff, N. Tintarev, K. Verspoor, M. C. Willemsen, and J. Zobel. Manifesto from Dagstuhl Perspectives Workshop 17442 – From Evaluating to Forecasting Performance: How to Turn Information Retrieval, Natural Language Processing and Recommender Systems into Predictive Sciences. *Dagstuhl Manifestos, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Germany*, 7(1):96–139, 2018.

[14] N. Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum*, 51(3):32–41, December 2017.

[15] D. J. Hand. *Measurement Theory and Practice: The World Through Quantification*. John Wiley & Sons, USA, 2010.

[16] O. Hölder. Die Axiome der Quantität und die Lehre vom Mass. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physikalische Classe*, 53:1–64, 1901.

[17] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement. Additive and Polynomial Representations*, volume 1. Academic Press, New York, USA, 1971.
[18] F. M. Lord. On the Statistical Treatment of Football Numbers. *American Psychologist*, 8(12):750–751, 1953.

[19] R. D. Luce, D. H. Krantz, P. Suppes, and A. Tversky. *Foundations of Measurement. Representation, Axiomatization, and Invariance*, volume 3. Academic Press, New York, USA, 1990.

[20] A. Moffat. Batch Evaluation Metrics in Information Retrieval: Measures, Scales, and Meaning. *IEEE Access*, 10:105564–105577, 2022.

[21] L. Narens. *Theories of Meaningfulness*. Lawrence Erlbaum Associates, Mahwah (NJ), USA, 2002.

[22] G. B. Rossi. *Measurement and Probability. A Probabilistic Theory of Measurement with Applications*. Springer-Verlag, New York, USA, 2014.

[23] T. Sakai. On Fuhr’s Guideline for IR Evaluation. *SIGIR Forum*, 54(1):p14:1–p14:8, June 2020.

[24] A. Z. Scholten and D. Borsboom. A reanalysis of Lord’s statistical treatment of football numbers. *Journal of Mathematical Psychology*, 53(2):69–75, April 2009.

[25] S. S. Stevens. On the Theory of Scales of Measurement. *Science, New Series*, 103(2684):677–680, June 1946.

[26] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky. *Foundations of Measurement. Geometrical, Threshold, and Probabilistic Representations*, volume 2. Academic Press, New York, USA, 1989.

[27] C. J. van Rijsbergen. Foundations of Evaluation. *Journal of Documentation*, 30(4):365–373, 1974.