IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 30, 2021, accepted September 20, 2021, date of publication September 29, 2021, date of current version October 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116857

Towards Meaningful Statements in IR Evaluation: Mapping Evaluation Measures to Interval Scales

MARCO FERRANTE[1], NICOLA FERRO[2], (Senior Member, IEEE), AND NORBERT FUHR[3]

[1]Department of Mathematics "Tullio Levi-Civita," University of Padua, 35122 Padua, Italy
[2]Department of Information Engineering, University of Padua, 35131 Padua, Italy
[3]Faculty of Engineering, University of Duisburg–Essen, 47057 Duisburg, Germany

Corresponding author: Nicola Ferro (ferro@dei.unipd.it)

ABSTRACT Information Retrieval (IR) is a discipline deeply rooted in evaluation since its inception. Indeed, experimentally measuring and statistically validating the performance of IR systems are the only possible ways to compare systems and understand which are better than others and, ultimately, more effective and useful for end-users. Since the seminal paper by Stevens (1946), it is known that the properties of the measurement scales determine the operations you should or should not perform with values from those scales. For example, Stevens suggested that you can compute means and variances only when you are working with, at least, interval scales. It was recently shown that the most popular evaluation measures in IR are not interval-scaled. However, so far, there has been little or no investigation in IR on the impact and consequences of departing from scale assumptions. Taken to the extremes, it might even mean that decades of experimental IR research used potentially improper methods, which may have produced results needing further validation. However, it was unclear if and to what extent these findings apply to actual evaluations; this opened a debate in the community with researchers standing on opposite positions about whether this should be considered an issue (or not) and to what extent. In this paper, we first give an introduction to the representational measurement theory explaining why certain operations and significance tests are permissible only with scales of a certain level. For that, we introduce the notion of *meaningfulness* specifying the conditions under which the truth (or falsity) of a statement is invariant under permissible transformations of a scale. Furthermore, we show how the recall base and the length of the run may make comparison and aggregation across topics problematic. Then we propose a straightforward and powerful approach for turning an evaluation measure into an interval scale, and describe an experimental evaluation of the differences between the original measures and the interval-scaled ones. For all the regarded measures – namely Precision, Recall, Average Precision, (Normalized) Discounted Cumulative Gain, Rank-Biased Precision and Reciprocal Rank - we observe substantial effects, both on the order of average values and on the outcome of significance tests. For the latter, previously significant differences turn out to be insignificant, while insignificant ones become significant. The effect varies remarkably between the tests considered but on average, we observed a 25% change in the decision about which systems are significantly different and which are not. These experimental findings further support the idea that measurement scales matter and that departing from their assumptions has an impact. This not only suggests that, to the extent possible, it would be better to comply with such assumptions but it also urges us to clearly indicate when we depart from such assumptions and, carefully, point out the limitations of the conclusions we draw and under which conditions they are drawn.

INDEX TERMS Information retrieval, measurement, software metrics, statistical analysis, experimental evaluation, retrieval effectiveness, interval scale.

I. INTRODUCTION

By virtue or by necessity, Information Retrieval (IR) has always been deeply rooted in experimentation and evaluation has been a formidable driver of innovation and advancement in the field, as also witnessed by the success of the major evaluation initiatives – Text REtrieval Conference (TREC)[1] in the United States [56], Conference and Labs of the Evaluation Forum (CLEF)[2] in Europe [44], NII Testbeds and

The associate editor coordinating the review of this manuscript and approving it for publication was Young Jin Chun.

[1]https://trec.nist.gov/
[2]http://www.clef-initiative.eu/

Community for Information access Research (NTCIR)[3] in Japan and Asia [93], and Forum for Information Retrieval Evaluation (FIRE)[4] in India – not only from the scientific and technological point of view but also from the economic impact one [85].

Central to experimentation and evaluation is the notion to measure the performance of an IR system and to reliably understand which systems perform better than others. To this end, there is a rich set of IR literature discussing existing evaluation measures or introducing new ones as well as proposing frameworks to model them [25], [79]. The major goal is to quantify users' experience of retrieval quality for certain types of search behavior, like e.g. users stopping at the first relevant document, or after the first ten results. Most of the measures proposed are based on plausible arguments and often accompanied by experimental studies, also investigating how close they are to end-user experience and satisfaction [61], [127], [128]. However, less attention has been given to the theoretic bases of evaluation measures, leading to possible violations of important assumptions and affecting the *validity* of the scientific results based on them, especially the *internal validity*, meant as ''the ability to draw conclusions about causal relationships from the results of a study'' [32, p. 157]. A few years ago, Robertson [84] raised the question of which scales are used by IR evaluation measures, since they determine which operations make sense on the values of a measure, as originally proposed by Stevens [103]. Scales have increasing properties: a *nominal scale* allows for determination of equality and for the computation of the mode; an *ordinal scale* allows only for determination of greater or less and for the computation of medians and percentiles; an *interval scale* allows also for determination of equality of intervals or differences and for the computation of mean, standard deviation, and rank-order correlation; finally, a *ratio scale* allows also for the determination of equality of ratios and for the computation of coefficient of variation. Recently, Ferrante et al. [40], [41] have theoretically shown that some of the most known and used IR measures, like Average Precision (AP) or Discounted Cumulative Gain (DCG), are not interval-scales. As a consequence, according to Stevens's prescriptions, we should neither compute means, standard deviations and confidence intervals, nor perform significance tests that require an interval scale.

Let us illustrate the problem with the example of reciprocal rank, which focuses on the first rank containing a relevant document. Regarding ranks as ordinary scale poses no problem. However, if we want to map the ranks onto an interval scale, there is the question if there are any constraints that we should take into account (besides being strictly monotonic). Obviously, the choice of the mapping function will also affect the outcome of system comparisons (e.g. when regarding means). A naive approach might regard any strictly monotonic function as permissible, the choice of the transformation just defining different standpoints. However, this assumption contradicts the essential property of interval scales of being based on intervals of equal length. A popular example are temperature scales, where Anders Celsius and Daniel Gabriel Fahrenheit came up with different definitions for the interval size (and the zero point), but both defined proper interval scales. In reciprocal rank, in contrast, all intervals have different lengths, and so it cannot be regarded as interval scale.

Over the decades there has been much debate about Stevens's prescriptions [54], [68], [75], [113]. This debate has also spawn to the IR field where Fuhr [49] suggested strict adherence to Stevens's prescriptions, motivated by the theoretical guarantees on the validity of the drawn conclusions. On the opposite, Sakai [92] argued for a more lenient approach, motivated by the usefulness of some measures and the fact that in other research fields they do average and compute parametric significance tests even on ordinal scales. The debate further sparked off on whether it was legitimate to make ''prohibitions'' and whether all researchers should adhere to the same guidelines or not.

Our vision is that it is neither a matter of ''prohibitions'' nor whether you wish (or not) to stick with them. Also, it is not a matter of opposing strict and formal constraints to the fact that measures may capture well the user experience and thus they should be used anyway. Rather, it is now time for the IR field to accurately investigate and understand the scale properties of its evaluation measures and their implications on the validity of our experimental findings. For example, we should explore whether and to what extent violations of the assumptions behind the scales used to measure system performance impact on the conclusions about which systems are better than others or which systems should actually go in production, because deemed more performing or because of more stable and generalizable performance.

As a matter of fact, we are not aware of any experimental IR paper that regarded evaluation measures as ordinal scales, thus refraining from computing (and comparing) means; also, most papers using evaluation measures apply parametric tests which, in a strict sense, should be used only from interval scales onwards. This means that potentially improper methods have been applied, without any preliminary verification of the possible consequences or any limitations and caveats on the drawn conclusions. Independently from your stance in the above long-standing debate, the key question about IR experimental findings is: are we on the safe side or are we at risk? Are we in a situation like using a rubber band for measuring and comparing lengths? Are we facing a state of the affairs where decades of IR research might have left unnoticed potential shortcomings?

We neither have the answer to these questions nor we pretend to say the final word on them. Rather, our intention with this paper is to lay the foundations and set all the pieces needed to have the instruments to answer these questions and to stimulate the IR community to rigoursly and systematically

[3]http://research.nii.ac.jp/ntcir/
[4]http://fire.irsi.res.in/

discuss these issues on a common ground in order to reach shared conclusions.

Therefore, the contributions of the paper are as follows:

1) introduction to the *representational measurement theory* [65], [71], [105], clearly explaining why (or why not) certain operations and significance tests should be permissible on a given scale and presenting the different stances on this long-standing debate;
2) introduction to the notion of *meaningfulness* [37], [81], [83], i.e. the conditions under which the truth (or falsity) of a statement is invariant under permissible transformations of a scale. To the best of our knowledge, this concept has never been investigated or applied in IR but it is fundamental to the validity of the inferences we draw;
3) discussion and demonstration of further measurement issues, specific to IR and beyond the debate on permissible operations. In particular, we show how the recall base and the length of the run may make averaging across topics (or other forms of aggregate statistics) problematic, at best;
4) proposal of a straightforward and powerful approach for turning an evaluation measure into an interval scale, by transforming its values into their rank position. In this way, we provide a means for improving the meaningfulness and validity of our inferences, still preserving the different user models embedded by the various evaluation measures;
5) experimental evaluation of the differences between using the original measures and the interval-scaled ones, by relying on several TREC collections. For all the regarded measures – namely Precision, Recall, AP, DCG, Normalized Discounted Cumulative Gain (nDCG), Rank-Biased Precision (RBP), and Reciprocal Rank (RR) – we observe substantial effects, both on the order of average values and on the outcome of significance tests. For the latter, previously significant differences turn out to be insignificant, while insignificant ones become significant. The effect varies remarkably between the tests considered but on average, we observed a 25% change in decisions about what is significant and what is not.

The latter experimental findings further support the idea that measurement scales matter and that departing from their assumptions has an impact. This not only suggests that we should strive to comply with such assumptions, at least to the extent it is possible; but it also urges us to clearly indicate when, in experimentation, we depart from such assumptions and, as a consequence, carefully point out the limitations of the conclusions we draw and under which conditions they are drawn. Moreover, we underline how the concept of *meaningfulness* could represent the keystone to start a rigorous discussion on IR evaluation measures, to set criteria to explore the impact of departures from Stevens's prescriptions, and to understand how to validate and limit our experimental finding when we wish to depart from such assumptions.

The paper is organized as follows: Section II introduces the basic concepts about experimental evaluation in IR; Section III provides an overview of the representational theory of measurement, of the different types of scale, and the notion of meaningfulness. Section IV deeply discusses measurement and meaningfulness issues specific to IR. Section V briefly summarizes some related works. Section VI explains our methodology for transforming evaluation measures into interval scales. Section VII introduces the experimental setup while Section VIII discusses the results of the experiments. Finally, Section IX draws some conclusions and gives an outlook on future work.

II. EXPERIMENTAL EVALUATION IN IR

Information Retrieval (IR) is concerned with complex systems delivering a variety of key applications to industry and society: web search engines [21], (bio)medical search [80], expertise retrieval systems [11], intellectual property and patent search [72], enterprise search [18], and many others.

IR systems operate using a best match approach: in response to an often vague user query, they return a ranked list of documents ordered by the estimation of their relevance to that query. In this context *effectiveness*, meant as the ability of systems to retrieve and better rank relevant documents while at the same time suppressing the retrieval of not relevant ones, is the primary concern. Since there are no a-priori exact answers to a user query, experimental evaluation [55] based on effectiveness is the main driver of research and innovation in the field. Indeed, the measurement of system performances from the effectiveness point of view is basically the only means to determine the best approaches and to understand how to improve IR systems.

Experimental evaluation addresses a very wide spectrum of cases, ranging from system-oriented evaluation [95] to user-oriented evaluation [63]. In this paper, we focus on system-oriented evaluation which is performed according to the *Cranfield paradigm* [30].

Figure 1 summarizes the Cranfield paradigm which is based on experimental collections $\mathcal{C} = (D, T, RJ)$ where: a corpus of documents D represents the domain of interest; a set of topics T represents the user information needs; and human-made relevance judgments RJ are the "correct" answers, or ground-truth, determining, for each topic, the relevant documents. Relevance judgments are typically expressed as either binary relevance, i.e. relevant or not relevant, or as graded relevance [62], e.g. not relevant, partially relevant, highly relevant. The ranked result lists, i.e. the IR system outputs, are then scored with respect to the ground-truth using several evaluation measures [88].

The main goal of this experimental setup is to be able to compare the performance of different IR systems in a robust and repeatable way, as they are all scored with respect to the same experimental collection. Experimental collections and evaluation measures are *controlled variables*, since they are kept fixed during experimentation; IR systems are *independent variables*, since they are the object of experimentation,

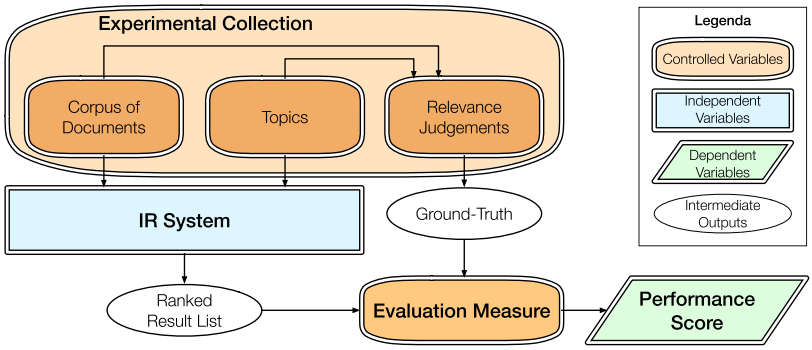


FIGURE 1. The Cranfield paradigm for experimental evaluation.

compared one against the other; and, performance scores are the *dependent variables*, since their observed value changes as IR systems change [48].

Carrying out experimental evaluations according to the Cranfield paradigm is very demanding in terms of both the time and the effort required to prepare the experimental collection. Therefore, it is usually carried out in publicly open and large-scale evaluation campaigns, often at international level, to share the effort, compare state-of-the-art systems and algorithms on a common and reproducible ground, and maximize the impact. Therefore, over the last 20 years, experimental evaluation has been carried out in large-scale evaluation campaigns at international level, such as TREC [56] in the US, CLEF [44] in Europe, NTCIR [93] in Japan and the other Asian countries, and FIRE in India.

In fact, IR evaluation adopts a whole breadth of evaluation measures [88] because different evaluation measures embed different user models in scanning the result list and thus represent different angles on the effectiveness of an IR system. Average Precision (AP) [17], Precision at Ten (P@10) [20], Rank-Biased Precision (RBP) [78], Normalized Discounted Cumulative Gain (nDCG) [60], and Expected Reciprocal Rank (ERR) [28] are among the most commonly adopted measures. Evaluation measures are typically studied in an empirical way, e.g. by using correlation analysis [119], discriminative power [86], [87], or robustness to pool downsampling [16], [124]. On the other hand, few studies have been undertaken to understand the formal properties of evaluation measures and they have just scratched the surface of the problem: [7], [13], [19], [39]–[41], [110].

Statistical analyses and statistical significance testing play a fundamental role in experimental evaluation [26], [59], [89], [97] since they provide us with the means to properly assess differences among compared systems and to understand when they actually matter.

Overall, despite IR being a discipline so strongly rooted in experimentation and being evaluation measures and statistical analyses so central to IR, to the best of our knowledge, there has been very limited attention and studies about how scale properties affect evaluation measures and statistical analyses in IR experimentation.

III. MEASUREMENT

A. OVERVIEW

The *representational theory of measurement* [65], [71], [105] is one of the most developed approaches to measurement, suitable for many areas of science ranging to physics and engineering to psychology. The basic idea is that real world *objects* have *attributes* which constitute their relevant features and induce a set of relationship among them; the set of objects E together with the relationships $R_1^E, R_2^E, \ldots$ among them comprise the so-called Empirical Relational System (ERS) $\mathbf{E} = \langle E, R_1^E, R_2^E, \ldots \rangle$. Then, we look for a mapping between the real word objects E and numbers N in such a way that the relationships $R_1^E, R_2^E, \ldots$ among the objects match with relationships $R_1^N, R_2^N, \ldots$ among numbers; the set of numbers N together with the relationships $R_1^N, R_2^N, \ldots$ constitutes the so-called Numerical Relational System (NRS) $\mathbf{N} = \langle N, R_1^N, R_2^N, \ldots \rangle$.

More precisely, the representational theory of measurement seeks for a *homomorphism* ϕ which maps E onto N in such a way that $\forall R_i^E, \forall e_1, e_2, \ldots, e_k \in E \mid (e_1, e_2, \ldots, e_k) \in R_i^E$ it holds that $\exists n_1 = \phi(e_1), n_2 = \phi(e_2), \ldots n_k = \phi(e_k) \in N \mid (n_1, n_2, \ldots, n_k) \in R_i^N$. The homomorphism ϕ is called a *scale of measurement*. Note that, in general, we seek for an homomorphism and not an isomorphism because two different real word objects might be mapped onto the same number.

The most typical example is length. Suppose the ERS $\mathbf{E} = \langle E, \succsim, \circ \rangle$ is a set of rods with an order relationship $\succsim$ among rods and a concatenation operation $\circ$ among them. If the attribute under examination is the length of a rod, we can map the ERS to the NRS $\mathbf{N} = \langle \mathbb{R}_0^+, \geq, + \rangle$ such that $\forall e_1, e_2, e_3 \in E$ it holds $e_1 \succsim e_2 \Leftrightarrow \phi(e_1) \geq \phi(e_2)$ and $e_1 \circ e_2 \sim e_3 \Leftrightarrow \phi(e_1) + \phi(e_2) = \phi(e_3)$, that is if a rod is longer than another one the number assigned to the first one is bigger than the number assigned to the second one and the concatenation of two rods corresponds to the sum of the two numbers assigned to them.

The core of the representational theory of measurements is to seek for a *representation* theorem and a *uniqueness* theorem for the scale of measurement in order to fully define it.

The *representation theorem* ensures that if the ERS satisfies given properties, it is possible to construct a homomorphism to a certain NRS. In the previous example, the representation theorem defines which properties the order relation $\succsim$ and the concatenation $\circ$ have to satisfy in order to construct a real-valued function ϕ which is order preserving and additive. It is important to underline that the representational theory of measurement seeks for "operations" among real word objects – e.g. we can put two rods side by side to order them or we can lay two rods end by end to concatenate them – and if these "operations" satisfy given properties they can be reflected into corresponding operations among numbers, where numbers are just a proxy of what happens among real world objects but are much more convenient to manipulate.

In general, given an ERS and an NRS, it is possible to create more than one homomorphism between them. For example, it is possible to express length by using meters or yards and both of them are legitimate scales for length. The *uniqueness theorem* is concerned with determining which are the *permissible transformations* $\phi \rightarrow \phi'$ such that ϕ and ϕ' are *both* homomorphisms of the given ERS into the *same* NRS. In our example, any transformation $\phi' = \alpha\phi, \alpha > 0$ is permissible for length. Therefore, the uniqueness theorem guarantees that the "structure" of a scale of measurement is invariant to changes in the numerical assignment, which preserve the relationships.

B. CLASSIFICATION OF THE SCALES OF MEASUREMENT

Stevens [103] introduced a classification of scales based on their permissible transformations, described below.

1) 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. The ERS consists only of different classes without any notion of ordering among them and 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.

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

Example 1 (Nominal Scale): Consider a classification of people by their country, e.g. France, Germany, Greece, Italy, Spain, and so on. We could define the two following measurements:

$$\phi = \begin{cases} 5 & \textit{if France} \\ 4 & \textit{if Germany} \\ 3 & \textit{if Greece} \\ 2 & \textit{if Italy} \\ 1 & \textit{if Spain} \\ \cdots & \textit{if } \cdots \end{cases} \qquad \phi' = \begin{cases} 41 & \textit{if France} \\ 13 & \textit{if Germany} \\ -10 & \textit{if Greece} \\ 23 & \textit{if Italy} \\ 17 & \textit{if Spain} \\ \cdots & \textit{if } \cdots \end{cases}$$

both ϕ and ϕ' are valid measures, which can be related with a one-to-one mapping. Note that even if ϕ looks like being ordered, there is actually no meaning in the associated magnitudes and so it should not be confused with an ordinal scale. Moreover, even if it is alway possible to operate with numbers, using ϕ and performing $4 - 3 = 1$, which would correspond to Germany $-$ Greece $\stackrel{?}{=}$ Spain, has no specific meaning, as well as using ϕ' and performing $13 - (-10) = 23$, which would correspond to Germany $-$ Greece $\stackrel{?}{=}$ Italy, even in disagreement with the previous case.

2) 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. The ERS consists of classes that are ordered with respect to the attribute under examination and 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. As a consequence, addition, subtraction or other mathematical operations have no meaning.

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

Example 2 (Ordinal Scale): The European Commission Regulation 607/2009 [35] and the follow-up regulation 2019/33 [36] set the following increasing scale to classify sparkling wines on the basis of their sugar content:

- pas dosé *(brut nature): sugar content is less than 3 grams per litre; let us call this range* $s_0 = [0, 3]$*;*
- extra brut*: sugar content is between 0 and 6 grams per litre; let us call this range* $s_1 = [0, 6]$*;*
- brut*: sugar content is less than 12 grams per litre; let us call this range* $s_2 = [0, 12]$*;*
- extra dry*: sugar content is between 12 and 17 grams per litre; let us call this range* $s_3 = (12, 17]$*;*
- sec *(dry): sugar content is between 17 and 32 grams per litre; let us call this range* $s_4 = (17, 32]$*;*
- demi-sec *(medium dry): sugar content is between 32 and 50 grams per litre; let us call this range* $s_5 = (32, 50]$*;*
- doux *(sweet): sugar content is greater than 50 grams per litre; let us call this range* $s_6 = (50, 2000]$*, where 2000 grams per litre is roughly the saturation of sugar in water, which is much higher than that of sugar in alcohol.*

We can introduce two alternative ordinal scales ϕ and ϕ' of the above wine scale where ϕ is given by the maximum of a range[5] *while ϕ' is given by a monotonic*

[5]Note that the EU regulations define not disjoint intervals and we stayed with the official definitions. This does not influence the definition of the ordinal scales above, since they are based on the maximum of each interval, which is unique and strictly increasing for each of them.

transformation $\phi' = \phi^2$*:*

$$\phi = \begin{cases} 3 & \textit{if pas dosé} \\ 6 & \textit{if extra brut} \\ 12 & \textit{if brut} \\ 17 & \textit{if extra dry} \\ 32 & \textit{if sec} \\ 50 & \textit{if demi-sec} \\ 2000 & \textit{if doux} \end{cases} \quad \phi' = \begin{cases} 9 & \textit{if pas dosé} \\ 36 & \textit{if extra brut} \\ 144 & \textit{if brut} \\ 289 & \textit{if extra dry} \\ 1024 & \textit{if sec} \\ 2500 & \textit{if demi-sec} \\ 4000000 & \textit{if doux} \end{cases}$$

As in the case of the previous Example 1, mathematical operations have no specific meaning, even if, especially in the case of ϕ*, we may be tempted to perform operations like* $\frac{brut}{extra\ brut} = \frac{12}{6} = 2$ *to express statements like "brut may be twice as sweet as extra brut". However, such a statement cannot be expressed on the* ϕ *or* ϕ' *scale and it actually comes from implicitly changing scale to the* mass concentration *scale of the solution, which is a ratio scale (see below) where the division operation would make sense. Also addition and subtraction have no meaning, so brut*$-$*extra brut* $= 12-6 = 6$ *is not a way to express statements like "brut may have* $6\ g/l$ *of sugar more than extra brut", for the same reasons above. We could perform operations such as* $\text{sgn}(\phi(e_1) - \phi(e_2))$ *or* $\text{sgn}(\phi'(e_1) - \phi'(e_2))$ *but this would be just a more involute way of expressing the order among categories, which is the only property guaranteed by ordinal scales.*

3) INTERVAL SCALE

Besides relying on ordered classes, it also captures information about the size of the intervals that separate the classes. The ERS consists of classes that are ordered with respect to the attribute under examination and where the *size of the "gap"* among two classes is somehow understood; more precisely, fundamental to the definition of an interval scale is that *intervals must be equi-spaced*. An interval scale preserves order, as an ordinal one, and differences among classes have meaning – but not their ratio. Therefore, addition and subtraction are acceptable operations but not multiplication and division.

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

Note that while ratios of classes $\frac{\phi(e_1)}{\phi(e_2)}$ have no meaning on an interval scale, the ratio of differences among classes, i.e. the ratio of intervals, is allowed and invariant $\frac{\phi'(a)-\phi'(b)}{\phi'(c)-\phi'(d)} = \frac{[\alpha\phi(a)+\beta]-[\alpha\phi(b)+\beta]}{[\alpha\phi(c)+\beta]-[\alpha\phi(d)+\beta]} = \frac{\phi(a)-\phi(b)}{\phi(c)-\phi(d)}$.

Example 3 (Interval Scale): A typical example of interval scale is temperature, which can be expressed on either the Fahrenheit or the Celsius scale, where the affine transformation $F = \frac{9}{5}C + 32$ *allows us to pass from one to the other. When talking about temperature it does not make sense to say that* $20\,^\circ C$ *is twice as hot as* $10\,^\circ C$*, i.e. multiplication and division are not allowed; you can also note that the division operation is not invariant to the transformation, since* $\frac{20\,^\circ C}{10\,^\circ C} = 2$ *but* $\frac{68\,^\circ F}{50\,^\circ F} = 1.36$*. However, it makes sense to say that the increase between* $10\,^\circ C$ *and* $20\,^\circ C$ *is the same as the increase between* $20\,^\circ C$ *and* $30\,^\circ C$*, i.e. addition and subtractions are allowed; you can also note that the subtraction operation is invariant to the transformation since* $30\,^\circ C - 20\,^\circ C = 20\,^\circ C - 10\,^\circ C = 10\,^\circ C$ *and* $86\,^\circ F - 68\,^\circ F = 68\,^\circ F - 50\,^\circ F = 18\,^\circ F$*. Moreover, the ratio of intervals* $\frac{20\,^\circ C - 10\,^\circ C}{30\,^\circ C - 20\,^\circ C} = 1$ *is invariant to the transformation* $\frac{68\,^\circ F - 50\,^\circ F}{86\,^\circ F - 68\,^\circ F} = 1$*.*

Central to the notion of temperature is the fact that the size of the "gap" has the same meaning all over the scale; indeed, 1 degree represents the same amount of thermal energy all over the scale. i.e. the gaps are equi-spaced*.*

4) RATIO SCALE

It allows us to compute ratios among the different classes. The ERS consists of classes that are ordered, where there is a notion of ''gap'' among two classes and where the ''proportion'' among two classes is somehow understood. It preserves order and differences as well as ratios. Therefore, all the arithmetic operations are allowed.

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

Example 4 (Ratio Scale): A typical example of a ratio scale is length which can be expressed on different scales, e.g. meters or yards, which can all be mapped one into another via a similarity transformation. For example, to pass from kilometers (ϕ) *to miles* (ϕ')*, we have the following transformation* $\phi' = 0.62\phi$*.*

Another example of ratio scale is the absolute temperature on the Kelvin scale where there is a zero element, which represents the absence of any thermal motion.

C. ADMISSIBLE STATISTICAL OPERATIONS

Stevens moved a step forward and linked the notion of scale with that of admissible statistical operations which can be carried out with that scale:

- *Nominal Scale:* The only allowable operation is counting number of items in each class, that is, in statistical terms, mode and frequency.
- *Ordinal Scale:* Besides the operations already allowed for nominal scales, median, quantiles, and percentiles are appropriate, since there is a notion of ordering.
- *Interval Scale:* Besides the operations allowed for ordinal scales, mean and standard deviation are allowable since they depend just on sum and subtraction.[6]
- *Ratio Scale:* Besides the operations allowed for interval scales, geometric and harmonic mean, as well as coefficient of variation, are allowable since they depend on multiplication and division.

These prescriptions originated several debates over the decades. Lord [68, p. 751] argued that ''since the numbers don't remember where they come from, they always

[6]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.

behave the same way, regardless'' and so any operation should be allowed even on ''football numbers'', i.e. a nominal scale; Gaito [51] reinforced this argument by distinguishing between the realm of the measurement theory, where Stevens's restrictions should apply, and the realm of the statistical theory, where these restrictions should not be applied, since other assumptions, such as normal distribution of the data, are those actually needed. Townsend and Ashby [107] replied back showing cases where performing operations inadmissible for a given scale of measurement may mislead the conclusions drawn by statistical tests. O'Brien [82] discussed the type of errors introduced when using ordinal data for representing an underlying continuous variable, classifying them into pure transformation errors, pure categorization errors, pure grouping errors, and random measurement errors. Velleman and Wilkinson [113] summarized the previous debate and argumented that once you are in the numerical realm every operation is admissible among numbers. Recently, Scholten and Borsboom [98] made a case of flaws in the original Lord's argument and, as a striking consequence, Lord's experiment would not be a counterargument to Stevens's restrictions but it would rather comply with them. In a very recent textbook, Sauro and Lewis [96] firmly supported Lord's view, at least in the case of ordinal scales, but with the caveat to not make claims on the outcomes of a statistical test that violates the underlying scale. So, for example, if you are on an ordinal scale and you detected a significant effect using a test which would require a ratio scale, you should not claim that that effect is twice as big as another effect but just that it is significant.

D. MEANINGFULNESS

The above observation brings the debate back to the core issue of what we should pay attention to. Indeed, both Hand [54] and Michel [75], [76] argued that the problem is not what operations you can perform with numbers but what kind of *inference* you wish to make from those operations and how much such inference has to be indicative of what actually happens among real world objects. Already Adams et al. [2, pp. 99-100] explicitly stated that

> Statistical operations on measurements of a given scale are not appropriate or inappropriate *per se* but only relative to the kinds of statements made about them. 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.

These statements opened the way for the development of the theory of *meaningfulness* [37], [81], [83], which is a central concept to clearly shape and define the questions discussed above: according to the adopted measurement scales, what processing, manipulation, and analyses can be conducted and what can we tell about the conclusions drawn from such processing?

Note that the statement ''A mouse weights more than an elephant'' is meaningful even if it is clearly false; indeed, its truth value, i.e. false, does not change whatever weight scale you use (kilograms, pounds, and so on). Therefore, as anticipated above, meaningfulness is a distinct concept from the one of truth of a statement and it is somehow close to the notion of invariance in geometry, since the truth value of a statement stays the same independently from the permissible scales used to express it.

Finally, note that the theory of meaningfulness has been developed in a fully formal way, using algebra and the axiomatic set theory. Even if this is well beyond the purpose of the present paper, which is just moving the first steps in this direction, it opens up for a future and more formal investigation of these concepts applied to the case of IR evaluation measures.

Example 5 (Meaningfulness for a Nominal Scale – Example 1 Continued): Suppose that we observe a set of 10 people, where 5 people are Spanish, 3 German, 1 Greek, and 1 Italian. According to ϕ *we would have* $P = [1\ 1\ 1\ 1\ 1\ 4\ 4\ 4\ 3\ 2]$ *while according to* ϕ' *we would have* $P' = [17\ 17\ 17\ 17\ 17\ 13\ 13\ 13\ -10\ 23]$. *In both cases, the statement "Most people come from Spain" is meaningful since, if we compute the mode of the values, it is* 1 *in the case of* ϕ *and* 17 *in the case of* ϕ' *which both correspond to Spain. On the other hand, the statement "The lowest quartile consist of Spanish people" is not meaningful, since it is true with* 1 *corresponding to Spain in the case of* ϕ *but is is false with* 13 *corresponding to Germany in the case of* ϕ'. *Indeed, the first statement about the mode involves just counting, which is an allowable operation for a nominal scale, while the second statement about the lowest quartile requires a notion of ordering not present in a nominal scale.*

Example 6 (Meaningfulness for an Ordinal Scale – Example 2 Continued): Suppose that we have two wineries X and Y. The first winery W_X *produced five bottles as follows:* extra brut, extra brut, brut, extra dry, *and* sec*; the second one* W_Y *produced five bottles as follows:* pas dosé, pas dosé, pas dosé, brut, *and* demi-sec. *Therefore, according to the scale* ϕ, *we have* $\phi(W_X) = [6\ 6\ 12\ 17\ 32]$ *and* $\phi(W_Y) = [3\ 3\ 3\ 12\ 50]$*; while according to the scale* ϕ', *we have* $\phi'(W_X) = [36\ 36\ 144\ 289\ 1024]$ *and* $\phi'(W_Y) = [9\ 9\ 9\ 144\ 2500]$. *The statement "The median of the first winery is greater than the one of the second winery" is meaningful since* $12 > 3$ *according to* ϕ *is true as well as* $144 > 9$ *according to* ϕ'*; so we could safely say that the first winery produces a little more* brut*-like wines than the second one, focusing on a more standard product. On the other hand, the statement "The average of the first winery is greater than the one of the second winery" is not meaningful since* $14.6 > 14.2$ *according to* ϕ *is true but* $305.8 > 534.2$ *according to* ϕ' *is false, which would lead us to draw basically opposite conclusions based on the scale we use. Indeed, the first statement about the median involves just the notion of ordering which is allowable on an ordinal scale,*

while the second statement about the average requires to sum values, which is not an allowable operation.

Example 7 (Meaningfulness for an Interval Scale – Example 3 Continued): 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 °*C and in Oslo is* 10 °*C while one month ago it was* 12 °*C and* 7 °*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: $T_P^C = [2\ 2\ 4\ 8\ 36]$ *and* $T_R^C = [1\ 2\ 4\ 15\ 34]$ *in Celsius degrees and, the same,* $T_P^F = [35.6\ 35.6\ 39.2\ 46.4\ 96.8]$ *and* $T_R^F = [33.8\ 35.6\ 39.2\ 59.0\ 93.2]$ *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 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 invariant to affine transformations. Indeed, let* $X = \{x_1, x_2, \ldots, x_n\}$ *and* $Y = \{y_1, y_2, \ldots, y_n\}$ *be two set of values on an interval scale; it holds that*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\phi'(x_i) > \frac{1}{n}\sum_{i=1}^{n}\phi'(y_i)\\
&\quad\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left[\alpha\phi(x_i)+\beta\right] > \frac{1}{n}\sum_{i=1}^{n}\left[\alpha\phi(y_i)+\beta\right]\\
&\quad\Leftrightarrow \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)+\beta > \alpha\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i)\right)+\beta\\
&\quad\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) > \frac{1}{n}\sum_{i=1}^{n}\phi(y_i)
\end{aligned}$$

Therefore, the statement "The mean of X is greater than the mean of Y" is always meaningful.

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.*

Also note that we may be tempted to compare the results of the arithmetic mean with those of the geometric mean to gain "more insights". For example, we might observe that the arithmetic mean in Paris is less than in Rome – $10.4 < 11.2$ *in Celsius degrees – but the opposite is true when we consider the geometric mean –* $5.40 > 5.27$ *in Celsius degrees. We might thus highlight that this due to the fact that the first (and lowest) value* 2 *in Paris is double than* 1 *in Rome and that the geometric mean rewards gains on lowest values; on the other hand, the arithmetic mean rewards gains on higher values and thus* 8 *in Paris is (almost) half than* 15 *in Rome and it contributes less. However, while the explanation why the geometric mean may differ from the arithmetic one is surely credible, the issue here is that the geometric mean could not be relied upon, as well as conclusions drawn from it, since it is based on operations not allowed for interval scales; indeed, if we consider exactly the same temperatures just on the Fahrenheit scale, we would reach opposite conclusions.*

Example 8 (Meaningfulness for a Ratio Scale – Example 4 Continued): If the air distance between Rome and Padua is (about) 400 *kilometers and the air distance between Rome and Oslo is (about)* 2,000 *kilometers, the statement "Rome and Oslo are five times as distant as Rome and Padua" is meaningful, even expressed in miles, since* $248.54 \backsim 5 \cdot 1{,}242.74$*.*

On the Kelvin scale for temperature, it does make sense to say that a thing is half as hot as another thing if, for example, the first one is 273 *K (almost* 0 °*C,* 32 °*F) and the second one is* 546 *K (almost* 273 °*C,* 523.4 °*F); you can note, however, how this statement does not hold if we consider Celsius and Fahrenheit degrees, since* $\frac{0}{273} = 0$ *while* $\frac{32}{523.4} = 0.06$*.*

Finally, let us show that the statement "The geometric mean of X is greater than the geometric mean of Y" is always meaningful. Indeed, let $X = \{x_1, x_2, \ldots, x_n\}$ *and* $Y = \{y_1, y_2, \ldots, y_n\}$ *be two sets of values on a ratio scale; it holds that*

$$\begin{aligned}
&\sqrt[n]{\prod_{i=1}^{n}\phi'(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\phi'(y_i)}\\
&\quad\Leftrightarrow \sqrt[n]{\prod_{i=1}^{n}\alpha\phi(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\alpha\phi(y_i)}\\
&\quad\Leftrightarrow \alpha\sqrt[n]{\prod_{i=1}^{n}\phi(x_i)} > \alpha\sqrt[n]{\prod_{i=1}^{n}\phi(y_i)}\\
&\quad\Leftrightarrow \sqrt[n]{\prod_{i=1}^{n}\phi(x_i)} > \sqrt[n]{\prod_{i=1}^{n}\phi(y_i)}
\end{aligned}$$

E. STATISTICAL SIGNIFICANCE TESTING

Siegel [101] and Senders [100] have discussed the implications of Stevens' classification and permissible operations in the case of statistical inference and parametric and non-parametric statistical significance tests. Here we consider the following tests:

- *Sign Test [53]:* Is a non parametric test which looks at the signs of the differences among two paired samples x_i and y_i; the null hypothesis is that the median of the differences is zero.
  The sign test requires samples to be on an *ordinal scale*, since it needs to determine the sign of their difference

or, equivalently, which one is greater. Note that the sign test discards the tied samples, i.e. when $x_i = y_i$.

- *Wilcoxon Rank Sum Test (or Mann-Whitney U Test) [53], [123]:* Is a non parametric test which looks at the ranks of two paired samples x_i and y_i; the null hypothesis is that the two samples have the same median.
  The Wilcoxon rank sum test requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.
- *Wilcoxon Signed Rank Test [53], [123]:* Is a non parametric test which looks at the signs and ranks of the differences among two paired samples x_i and y_i; the null hypothesis is that the median of the differences is zero.
  The Wilcoxon signed rank test requires samples to be on an *interval scale*, since it regards the ranks of the differences, for which intervals must be equi-spaced. Note that the Wilcoxon signed rank test discards the tied samples, i.e. when $x_i = y_i$.
- *Student's t Test [104]:* Is a parametric test for the null hypothesis that two paired samples x_i and y_i come from a normal distribution with same mean and unknown variance.
  The Student's t test requires samples to be on an *interval scale*, since it needs to compute means and variances.
- *ANOVA [45], [67]:* Is a parametric test for the null hypothesis that q samples come from a normal distribution with same mean and unknown variance.
  ANalysis Of VAriance (ANOVA) requires samples to be on an *interval scale*, since it needs to compute means and variances.
- *Kruskal-Wallis Test [53], [66]:* Is a nonparametric version of the one-way ANOVA for the null hypothesis that q samples come from a distribution with the same median. It is based on the ranks of the different samples and it can be considered as an extension of the Wilcoxon rank sum test to the comparison of multiple systems at the same time.
  The Kruskal-Wallis test requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.
- *Friedman Test [46], [47], [53]:* Is a nonparametric version of the two-way ANOVA for the null hypothesis that the effects of the q samples are the same. It is based on the ranks of the different samples.
  The Friedman test requires samples to be on an *ordinal scale*, since it needs to order them for determining their rank.

As in the case of Stevens' permissible operations, defining which statistical significance tests should be permitted on the basis of the scale properties of the investigated variables raised a lot of discussion and controversy. Anderson [10], along the line of reasoning of Lord, argued that statistical significance tests should be used regardless of scale limitations. Gardner [52] summarizes much of the discussion up to that point, leaning towards not worrying too much about scale assumptions, and suggests that, if and when lack of compliance to measurement scale requirements biases the outcomes of significance tests, transformations can be applied to turn ordinal scales into more interval-like ones such as, for example, averaging the ranks of each score, as proposed by Gaito [50], or using a more complex set of rules, as developed by Abelson and Tukey [1]. Ware and Benson [121] replied to Gardner's positions by further revising the pro and con arguments and concluding that parametric significance tests should be used only when dealing with interval and ratio scales while, in the case of ordinal scales, nonparametric significance tests should be adopted. Townsend and Ashby [107] further investigated the issue, highlighting some serious pitfalls you may fall in, when ignoring the scale assumptions.

We can summarise the discussion with the conclusions of Marcus-Roberts and Roberts [74, p. 391]:

> The appropriateness of a statistical test of a hypothesis is just a matter of whether the population and sampling procedure satisfy the appropriate statistical model, and is not influenced by the properties of the measurement scale used. However, if we want to draw conclusions about a population which say something basic about the population, rather than something which is an accident of the particular scale of measurement used, then we should only test meaningful hypotheses, and meaningfulness is determined by the properties of the measurement scale in connection with the distribution of the population.

and Hand [54, p. 471]

> Restrictions on statistical operations arising from scale type are more important in model fitting and hypothesis testing contexts than in model generation or hypothesis generation contexts.

IV. MEASUREMENT ISSUES IN INFORMATION RETRIEVAL

A. WHY DOES STUDYING THE SCALE PROPERTIES OF IR EVALUATION MEASURES MATTER?

Let us start our discussion by considering a not-exhaustive list of core IR areas where scales may matter.

The most common and basic operation we perform to understand whether a system A is better than a system B is to average their performance over a set of topics and compare these aggregate scores. According to the discussion so far this leads to *meaningful statements* only if IR evaluation measures are, at least, interval scales.

Topic difficulty [24] is another central theme in IR because of its importance for adapting the behaviour of a system to the topic at hand. Voorhees [116], [117], in the TREC Robust tracks, explored how to evaluate and compare systems designed to deal with difficult topics and proposed to use the geometric mean, instead of the arithmetic one, for Average Precision (AP) [17]. However, the use of a geometric mean further raises the requirements for the evaluation measures, even calling for a ratio scale.

Statistical significance testing has a long story of adoption and investigation in IR, from the early uses of t-test reported by Salton and Lesk [94], to the discussion on the compliance with the distribution assumptions of significance tests by van Rijsbergen [111], to advocating for a more wide-spread adoption of different types of significance tests by Hull [59], Savoy [97], Carterette [26], Sakai [89], to surveys on the current state of adoption of significance tests by Sakai [91]. Again, drawing *meaningful inference* depends on the appropriate use of parametric or nonparametric tests in accordance with the scale properties of the adopted IR evaluation measures.

Several authors have proposed the use of score transformation and standardisation techniques, such as z-score by Webber et al. [122] and other types of linear (and non-linear) transformations by Sakai [90], Urbano et al. [109], in order to compare performance across collections and to reduce the impact of few topics skewing the performance distribution. However, in order to ensure *meaningful conclusions* from these transformation, at least an interval scale would be required.

Despite the many aspects of IR evaluation which can be affected by the scale properties of evaluation measures and despite the deep scrutiny that the above techniques have received over the years, there has been much less attention to the implications of the scale assumptions on them.

Robertson [84] was the first to discuss the admissibility of the use of the geometric mean from the Stevens's perspective in the context of the TREC Robust track. In particular, Robertson focused on the fact that Mean Average Precision (MAP) and Geometric Mean Average Precision (GMAP) may lead to different conclusions – e.g. blind feedback is beneficial according to MAP but detrimental according to GMAP – and which of them may hold more *(intrinsic) validity*. In this respect, Robertson [84, p. 80] observed that

> If the interval assumption is not valid for the original measure nor for any specific transformation of it, then *any* monotonic transformation of the measure is *just as good a measure* as the untransformed version. If we believe that the interval assumption is good for the original measure, that would give the arithmetic mean some validity over and above the means of transformed versions. If, however, we believe that the interval assumption might be good for one of the transformed versions, we should perhaps favour the transformed version over the original. But if there is no particular reason to believe the interval assumption for any version, then all versions are equally valid. If they differ, it is because they measure different things.

Since both AP and the log-transformation of AP (implied by the geometric mean) are not interval scales, Robertson concluded that no preference could be granted to MAP or GMAP in terms of (intrinsic) validity of their findings. In this way Robertson takes a neutral stance with respect to the debate on whether certain operations should be permitted or not on the basis of the scale properties.

Note that Robertson somehow implicitly indicates transformations as a possible means to turn a not-interval scale into an interval one, as also supported by Gaito [50], Abelson and Tukey [1].

As a final remark, even if Robertson did not mention it explicitly, his reasoning seems to be loosely related to the concept of *meaningfulness* when he says [p. 80]

> Good robustness would be indicated if the conclusions looked the same whatever transformation we used; if we found it easy to find transformations which would substantially change the conclusions, then we might infer that our conclusions are sensitive to the interval assumption, and that the different transformations measure different things in ways that may be important to us

still keeping a neutral stance about what should or should not be done.

Fuhr [49] took a firm position and argued that Mean Reciprocal Rank (MRR) [102] should not be computed because: 1) in general, RR is just an ordinal scale and, according to Stevens means cannot be computed for a ordinal scales; 2) in particular, RR has some counter-intuitive behaviour. On the other hand, Sakai [92] has recently disagreed with Fuhr: 1) in general, on the fact that means should not be computed for an ordinal scale, using arguments similar to those discussed in Section III-C; 2) in particular, on the use of RR which Sakai finds quite useful from a practical point of view.

Whatever stance you wish to take about whether (or not) operations should be constrained by scale properties, from the discussion so far, it clearly emerges that IR needs further and systematic investigation about the implications and impact of derogating from compliance with scale properties. Moreover, most of the above discussion is just about averaging values and does not tackle the implications for statistical significance testing. Finally, and more importantly, we completely lack a thorough discussion on and any adoption of the notion of *meaningfulness* in IR. This is quite striking for a discipline so strongly rooted in experimentation and so much based on inference.

B. A FORMAL THEORY OF SCALE PROPERTIES FOR IR EVALUATION MEASURES

Ferrante et al. [39]–[41] leveraged the representational theory of measurement for developing a formal theory of IR evaluation measures which allows us to determine the scale properties of an evaluation measure. In particular, they defined an ERS for system runs and used two basic operations – *swap*, i.e. swapping a relevant with a not-relevant document in a ranking, and *replacement*, i.e. substituting a relevant document with a not-relevant one – to study how runs are ordered. In this way, they demonstrated that there exists a *partial order of runs* where, when runs are comparable, all the measures agree on the same way of ordering them; however, when runs are not comparable, measures may disagree on how to order

them. By using properties of the partial orders and theorems from the representational theory of measurement, they were able to define an interval scale measure ϕ and to check if there is any linear transformation between such a measure ϕ and IR evaluation measures, in order to determine if the latter are interval scales too.

In short, Ferrante et al. found that, for a single topic:

- set-based evaluation measures:
  - binary relevance: precision, recall, F-measure are interval scales;
  - multi-graded relevance: Generalized Precision (gP) and Generalized Recall (gR) are interval scales only if the relevance degrees are on a ratio scale;
- rank-based evaluation measures:
  - binary relevance: Rank-Biased Precision (RBP) [78] is an interval scale only for $p = 1/2$; Average Precision (AP) is not an interval scale;
  - multi-graded relevance: Graded Rank-Biased Precision (gRBP) is an interval scale only for $p = G/(G + 1)$, where G is the normalized smallest gap between the gain of two consecutive relevance degrees, and the relevance degrees themselves are on a ratio scale; Discounted Cumulative Gain (DCG) [60] and Expected Reciprocal Rank (ERR) [28] are not interval scales.

Ferrante et al. [41] also studied what they called the *induced total order*, i.e. pretending that runs in the ERS are ordered by the actual values of a measure. Also in this case which is the most "favourable" to each measure, Ferrante et al. have shown that AP, RBP with $p \neq 1/2$ (and its multi-graded version), DCG, and ERR are not interval scales, because their values are not equi-spaced.

Figure 2 shows the *Hasse diagram* [34] which represents the partial order among all the runs[7] of length $N = 4$. In the figure, vertices are runs while edges represent the direct predecessor relation that is, if $r \prec s$, i.e. r and s are comparable, then r is below s in the diagram. Note that if r and s lie on the same horizontal level of the diagram, then they are incomparable; furthermore, elements on different levels may be incomparable as well. In the example $(1, 1, 0, 1) \preceq (1, 1, 1, 0)$, $(1, 1, 0, 0) \preceq (1, 1, 1, 0)$, and $(1, 0, 1, 1) \preceq (1, 1, 1, 0)$ are all comparable; therefore, all IR measures agree on these runs and order them in the same way. On the other hand, $(1, 1, 0, 0)$ and $(1, 0, 1, 1)$ are not comparable, as well as $(1, 1, 0, 0)$ and $(0, 1, 1, 1)$, and IR

[7]Note that, in general, we assume to operate in the finite and discrete case, i.e. finite number of documents, finite number of relevant documents, finite length of the run, since this reflects what happens in running systems.
Note also that, when constructing a scale, you need to account for all the possible cases so that we can use the same scale for all of them. For example, when constructing the length scale, you would consider people of all possible heights, even if in a specific regions you may observe only taller or shorter people. Therefore, to create a scale which holds for different topics and different collections of documents, we need to consider all the possible runs and all the combinations of relevant and not relevant documents. In this way, we can accommodate on the same scale topics without relevant documents, with just one relevant document, with two relevant documents, and so on up to many relevant documents.

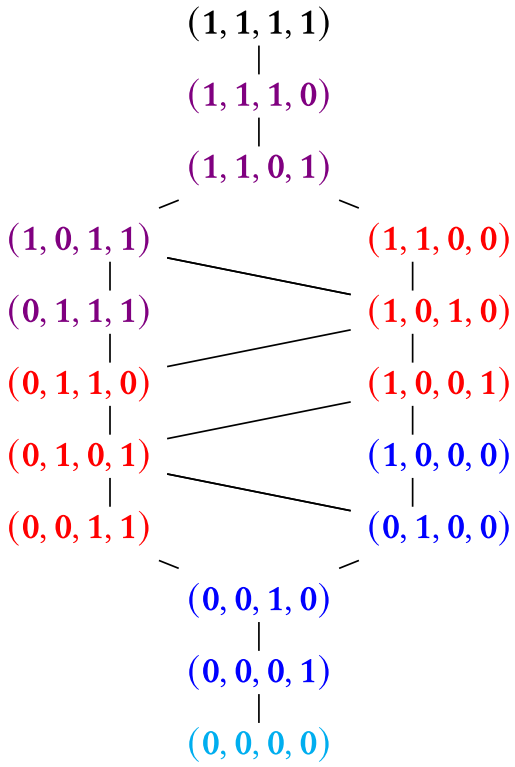
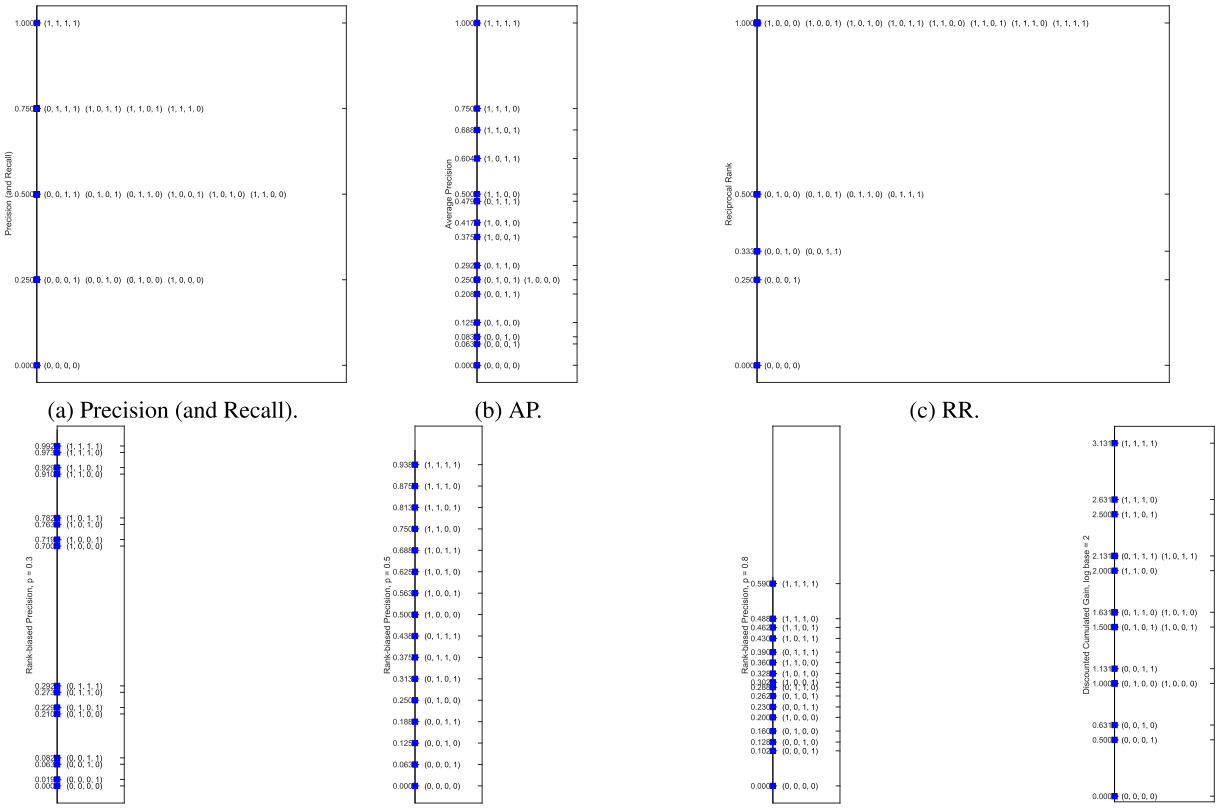


FIGURE 2. Hasse diagram showing the partial order of all the possible runs of length 4. The different colours of the runs correspond to different total numbers of relevant retrieved documents.

measures disagree on how to order them; as a consequence, measures will order these runs differently, producing different Rankings of Systems (RoS).

The difference in the RoS produced by evaluation measures is what is studied when performing a correlation analysis among measures, e.g. by using Kendall's τ [64]; practical wisdom says that measures should be neither too much correlated – otherwise it practically makes no difference using one or the other – nor too little correlated – otherwise it may be an indicator of some "pathological" behaviour of a measure. Indeed, each evaluation measure embodies a different user model [25], i.e. a different way in which the user interacts with the ranked result list and derives gain from the retrieved documents, and the differences between the RoS produced by different evaluation measures, and as a consequence their Kendall's τ may be considered as the tangible manifestation of such different user models. Note that the work by Ferrante et al. provides a formal explanation of the differences in Kendall's τ: for all the runs which are comparable in the Hasse diagram, Kendall's τ between different measures is 1, since all of them order these runs in the same way; for runs which are not comparable in the Hasse diagram, Kendall's τ between different measures is less than 1, since all of them order these runs differently; therefore, these not comparable runs are where user models differentiate themselves and can take a different stance.

However, these differences in the RoS are not causing IR evaluation measures to not be interval scales; they would



(a) Precision (and Recall). (b) AP. (c) RR.

(d) RBP, $p = 0.3$. (e) RBP, $p = 0.5$. (f) RBP, $p = 0.8$. (g) DCG, log base 2.

FIGURE 3. Ordering and spacing of the runs of Figure 2 by different evaluation measures. Each blue square corresponds to a score of a given measure. On the right of the square, the run corresponding to that score is reported; in case of tied runs, i.e. runs for which the measures produces the same score, they are all listed on the right of the square.

just mean that IR evaluation measures are different scales. The real problem with IR evaluation measures is that their scores are not *equi-spaced* and thus they cannot be interval scales, as explained in Section III-B. This issue is depicted in Figure 3 which shows how different measures – namely, Precision (and Recall[8]), AP, RR, RBP with $p \in \{0.3, 0.5, 0.8\}$, and DCG with log base 2 – order and space the runs shown in the Hasse diagram of Figure 2.

We can observe that only Precision (Recall) and RBP with $p = 0.5$ produce equi-spaced values, while all the other measures violate this assumption, required to obtain an interval scale; in other terms, Figure 3 visually represents the issue found by Ferrante et al. [41] even when using the induced total order. We can also note that all the measures agree only on the common comparable runs – i.e. $(0, 0, 0, 0) \preceq (0, 0, 0, 1) \preceq (0, 0, 1, 0)$ and $(1, 1, 0, 1) \preceq (1, 1, 1, 0) \preceq (1, 1, 1, 1)$ – but, as soon as incomparable runs come into play, they start to disagree on how to order them. Finally, looking at Figure 3 we can notice how IR measures behave differently in violating the equi-spacing assumption. RBP with $p \in \{0.3, 0.8\}$ and DCG follow a somehow regular pattern, where scores are not equi-spaced but they are in some way evenly clustered and they are symmetric if you fold the figure along its middle horizontal axis; on the other hand, AP and RR follow a much more irregular and not symmetric pattern.

We can also note how these measures spread values in their range differently. Precision (and Recall) and DCG spread their values all over the possible range while this is not always the case with RBP. Indeed, RBP assumes runs of infinite length and normalizes by the $\frac{1}{1-p}$ factor. However, we deal with runs of limited length and the $\frac{1}{1-p}$ factor is an overestimation, the bigger the overestimation the bigger is the value of p and the smaller is the length of the run – this is more clearly visible in the case of RBP with $p = 0.8$ in Figure 3f. Finally, AP, RBP with $p = 0.3$, and RR, i.e. those measures farther from being interval scales, leave large portions of their possible range completely unused. In particular, AP leaves one quarter of its range unused, in the top part roughly corresponding to the first quartile of the possible values; RR leaves one half of its range unused, in the top part roughly corresponding to the first and second quartiles of the possible values; and, finally, RBP with $p = 0.3$ leaves half of its range empty, in the middle part roughly corresponding to the second and third quartile of the possible values.

Why does it matter how much equi-spaced the values are and how they are spread over their range? Consider a random variable X that takes values in the set $\{0, 1, 2, 4, 13\}$. Even if

[8]Note that in this specific case, since the length of the run $N = 4$ and the recall base $RB = 4$ are the same, Precision and Recall yield to the same scores.

TABLE 1. Example for AP not being equi-spaced, considering the MAP of each systems across the queries Q_1 and Q_2.

System	Q_1	Q_2	MAP	System	Q_1	Q_2	MAP
A	$(1,1,1,0)$	$(1,0,0,1)$	0.5625	C	$(0,0,1,0)$	$(0,1,0,1)$	0.1667
B	$(1,1,0,1)$	$(1,0,1,0)$	0.5521	D	$(0,0,0,1)$	$(0,1,1,0)$	0.1771

all these five values can be obtained with equal probability, i.e. the random variable is uniform, the mean and the median of the variable differ, being the mean equal to 4 and the median to 2. This shows how the lack of equi-spacing causes some sort of ''imbalance'' even in the case of a uniform variable, which may be an undesirable situation from the measurement point of view, at least if not explicitly considered and accounted for. Furthermore, when we compute $\mathbb{P}[X \in (x-\varepsilon, x+\varepsilon)]$, i.e. the probability that the value of X is equal to x with an error of at most 2ε, this function is not constant all over the range but it is assumer greater for values around $\{0,1,2\}$ than for those around $\{4,13\}$. As a consequence, a similar accuracy in approximating the value of X produces a different precision in the measurement depending on the value x that we are considering. Note that in the present toy model it happens for $\varepsilon \geq 1$, but a suitable modification of the present model can produce the same behaviour for any $\varepsilon > 0$ set in advance.

As a further example, let us consider a measure with a limited range of equi-spaced values. If we draw a set of random values taken from this range and consider its arithmetic mean, by the law of large numbers, we have that this mean converges to the middle point of the range interval. This property is independent from the distance among the subsequent values, i.e. the unit of measurement chosen. So we can use such a procedure – the convergence of the mean towards the middle of the range – in order to ''calibrate'' the measuring instrument, independently from the specific unit of measurement chosen. This is no more possible if we have values which are not equi-spaced.

Example 9 (Effect of RR Not Being Equi-Spaced): Let us assume that we have two queries and two systems. System A returns the first relevant document at ranks 1 and 4, respectively, while system B finds the relevant answers in both cases at rank 2. Computing the MRR of the two systems, i.e. the average value of the RR, we get MRR(A) $= \frac{1}{2}(1/1 + 1/4) = 0.625$*, while MRR(B)* $= 0.5$*, telling us that system A is better than B. However, if instead of reciprocal rank, we regard the ranks themselves, we have equi-spaced values forming an interval scale (actually, even a ratio scale). In our example, system A finds the first relevant item on average at rank 2.5, which is worse than the average rank 2 of system B – so we would get the opposite finding when we use a scale still based on the rank of relevant documents but properly equi-spaced.*

Example 10 (Effect of AP Not Being Equi-Spaced): Table 1 shows an example of two system pairs (A,B) and (C,D) and two queries, for which we compute AP values. In the first case, mean of AP over the two queries, i.e. MAP, will say that A performs better than B, while in the second case, C is worse than D. Why is this effect related to AP not being on an interval scale? Because in both the examples, the runs retrieved by the two systems for a given topic have the same relevance degrees in the first two positions and just a swap of a relevant with a not-relevant document in the last two positions. So, to the same loss of relevance for the swap in the last two positions, still keeping the same relevance in the first two positions, AP ''reacts'' in one case telling us that system A is better than system B, and in the second case that D is better than C and this is also due to the not-equispaced values of AP, e. g. runs ranked 13 and 14 are much closer than runs ranked 10 (on the left branch of Figure 2) and 9, as shown in Figure 3b. Note that here we are neither questioning the top-heaviness of a measure nor its capability of reflecting user preferences but rather we point out how the lack of equi-spaced values affects the assessment supported by a measure.

The fact that IR evaluation measures, apart from Precision, Recall, and RBP with $p = 0.5$, are not interval scales leads to the general issues with computing means, statistical tests, and meaningfulness discussed in Sections from III-B to VIII-C and shown in Examples 3 and 7. In addition, Examples 9 and 10 above show how the lack of equi-spacing may also lead to statements like ''system A is better than B'' (or viceversa) which are not always intuitive all over the scale.

C. AVERAGING ACROSS TOPICS AND CORRELATION ANALYSIS REVISITED

The fact that Precision and Recall are interval scales makes addition and subtraction permissible operations and, as a consequence, computing arithmetic means permissible too. Therefore, it is safe to average performance of IR systems across topics when we use Precision and Recall. But is that really true?

As said, Ferrante et al. [41] have found an interval scale ϕ, called Set-Based Total Order (SBTO), and have shown that both Precision and Recall are an affine transformation of this interval scale and thus also an affine transformation of each other. Ferrante et al. [42] have raised this question: if Precision and Recall are transformations of the same interval scale, they are ordinal scales too and they should rank systems in the same way. Therefore, if they produce the same RoS, Kendall's τ correlation between them should be 1. So, why their Kendall's τ correlation is 0.8588, using the TREC 8 Ad-hoc data?

Let us consider how correlation analysis between evaluation measures works. Given two rankings X and Y, their Kendall's τ correlation is given by $\tau(X,Y) = \frac{P-Q}{\sqrt{(P+Q+T)(P+Q+U)}}$, where P is the total number of concordant pairs (pairs that are ranked in the same order in both vectors), Q the total number of discordant pairs (pairs that are ranked in opposite order in the two vectors), T and U are

the number of ties, respectively, in the first and in the second ranking. $\tau \in [-1, 1]$ where $\tau = 1$ indicates two perfectly concordant rankings, i.e. in the same order, $\tau = -1$ indicates two fully discordant rankings, i.e. in opposite order, and $\tau = 0$ means that 50% of the pairs are concordant and 50% discordant.

The typical way of performing correlation analysis is as follows: let ϕ_1 and ϕ_2 be two evaluation measures; in our case, ϕ_1 is Precision and ϕ_2 is Recall. Let Φ_1 and Φ_2 be two $T \times S$ matrices where each cell contains the performance on topic i of system j according to measures ϕ_1 and ϕ_2, respectively. Therefore, Φ_1 and Φ_2 represent the performance of S different systems (columns) over T topics (rows). Let $\overline{\Phi}_1$ and $\overline{\Phi}_2$ be the column-wise averages of the two matrices, i.e. the average of the performance of each system across the topics. If you sort systems by their score in $\overline{\Phi}_1$ and $\overline{\Phi}_2$, you obtain two RoS corresponding to ϕ_1 and ϕ_2, respectively, and you can compute Kendall's τ correlation between these two RoS. This is the traditional way for computing the correlation between two evaluation measures and Ferrante et al. call it *overall correlation*, since it first computes the average performance across the topics and then it computes the correlation between evaluation measures. This approach leads to a Kendall's τ correlation of 0.8588 between Precision and Recall.

Ferrante et al. proposed a different way of computing the correlation, called *topic-by-topic correlation*, where, for each topic i, they consider the RoS on that topic corresponding to ϕ_1 and the one corresponding to ϕ_2, i.e. they consider the i-th rows of Φ_1 and Φ_2, respectively; they then compute Kendall's τ correlation among the two RoS on that topic. Therefore, they end-up with a set of T correlation values, one for each topic. Using this way of computing correlation, Ferrante et al. found that Kendall's τ correlation between Precision and Recall is always 1 for all the topics and this was the result expected for two interval scales which order systems in the same way.

Therefore, if you consider each topic alone, Precision and Recall are just a transformation of the same interval scale, as Celsius and Fahrenheit are, and their Kendall's τ correlation is 1. However, if you first average across topics, which should be a permitted operation for interval scales, and then you compute Kendall's τ correlation, it stops to be 1. This was somehow surprising and unexpected. Indeed, as an example from another domain, if you take a matrix of scores in Celsius degrees and another one with the corresponding Fahrenheit degrees, their Kendall's τ correlation is always 1, either if you compute it row-by-row (i.e. our topic-by-topic correlation) or if you first average across rows and then compute it (i.e. our overall correlation).

Ferrante et al. [42, p. 305] explained this behaviour as due to the recall base:

> Recall heavily depends on the recall base which changes for each topic and it is used to normalize the score for each topic; therefore, in a sense, recall on each topic changes the way it orders systems

We further investigate this issue in Section IV-D below, where we provide details and demonstrations, but here we summarise our findings. The difference between overall and topic-by-topic correlation is basically due to the fact that we are using different interval scales for each topic. These scales are indeed transformations of one in the other for each topic and this is why topic-by-topic correlation is 1; however, since we are changing scale from one topic to another, when averaging across topics we are mixing different scales and this is why the overall correlation is different from 1.

Example 11 (Recall Corresponds to Different Scales on Different Topics): Let us consider Recall and let us assume that we have three queries q_1, q_2, q_3, with one, two and three relevant documents, respectively. Then, the possible values of Recall are as follows: for q_1 we have 0 and 1; for q_2 we have 0, $\frac{1}{2}$ and 1; and for q_3 we have 0, $\frac{1}{3}$, $\frac{2}{3}$, and 1. Obviously, we have three different interval scales here – although they are in the same range $[0, 1]$, their possible values are different. So we have to map the values onto a single scale, before we can do any statistics. There are two possibilities for doing this:

1) *We take the union of the possible values. This would yield the set $\{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$. However, these values are no longer equidistant, so it is not an interval scale.*
2) *We extend the union scale from above by additional values such that we have equidistant values, based on the least common denominator. Then we would have the set $\{0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, 1\}$ in our example. However, in this scale, the values $\frac{1}{6}$ and $\frac{5}{6}$ are not possible for our three example topics, and impossible values are not considered in the definition of the equidistance property of interval scales. Only if we had a fourth query with six relevant documents, this scale would be ok. In most cases, however, no such scale exists, and so the aggregated scale is not an interval one.*

The fact that we may be changing scale from topic to topic has very severe consequences. All the debate originated by Stevens's permissible operations and the possibility of averaging only from interval scales onwards has always been based on the obvious assumption that the averaged values were all drawn from the same scale; no one has ever doubted that it is not possible to average values coming from different scales because this would be like mixing apples with oranges. So, what is the meaningfulness of typical statements like "System A is (on average) better than system B" when we are not only violating the interval scale assumptions but, even more seriously, we are mixing different scales? What about the meaningfulness of typical statements like "System A is significantly better than system B"? The debate between using parametric or nonparametric tests concerns how much you wish to comply with the interval scale assumptions but, undoubtedly, all the significance tests, when aggregating across values, expect them to be drawn from the same scale.

If we wish to make an analogy, it is like the difference between using mass and weight, being Precision similar to mass and Recall to weight. It would be somehow safe to

average the mass of bodies coming from different planets but it would not to average their weight, due to the different gravity on the different planets. The recall base is what changes the gravity from planet/topic to planet/topic in the case of Recall.

However, even Precision is not completely ''safe'' because, when the length of the run changes, its scale changes as well. As a consequence we may end up using different scales from one run to another and this can happen not only across topics, as in the case of Recall, but also within topics, if we have two or more runs retrieving a different number of documents for that topic. This statements affects the evaluation of classical Boolean retrieval, where Precision and Recall are computed for the set of retrieved documents for each query, followed by averaging over all queries. So we have to conclude that this procedure is seriously flawed. Luckily, in most of today's evaluations, the length of the run has a much smaller effect because, in typical TREC settings, almost all the runs retrieve 1,000 documents for each topic and just few of them retrieve less documents; this effect would also (practically) disappear when you consider Precision at lower cut-offs, like P@10, when it is almost guaranteed that all the runs retrieve 10 documents.

Summing up, independently from an evaluation measures being an interval scale or not, the recall base (greatly) and the length of the run (less) cause the scale to change from topic to topic and/or from run to run. This makes averaging across topics, as well as other forms of aggregation used in significance tests, problematic at best. We show how and why this happens in the case of Precision (Section IV-D1) and Recall (Section IV-D2), which are the simplest measures you can think of, since they change either the length of the run or the recall base alone. We also consider the more complex case of the F-measure (Section IV-D3), which changes both the run length and the recall base at the same time. Therefore, we hypothesise that these issues may be even more severe in the case of more complex evaluation measures, like AP and others, which are not even interval scales and mix recall base and run length with rank position and various forms of utility accumulation and stopping behaviours.

Finally, also the way in which we interpret the results of correlation analysis may be impacted. Indeed, we typically attribute differences in correlation values to the different user models embedded by evaluation measures. The rule-of-thumb by Voorhees [114], [115] is that an overall correlation above 0.9 means that two evaluation measures are practically equivalent, an overall correlation between 0.8 and 0.9 means that two measures are similar, while dropping below 0.8 indicates that measures are departing more and more. Therefore a correlation of 0.8588 would suggest that Precision and Recall share some commonalities but they differ enough due to their user models, still not being pathologically different. However, we (now) know that they are just the transformation of the same interval scale and that this correlation value is just an artifact of mixing different scales across topics rather than an intrinsic difference in the user models of Precision and Recall.

D. WHY MAY SCALES CHANGE FROM TOPIC TO TOPIC OR FROM RUN LENGTH TO RUN LENGTH?

As discussed above, Ferrante et al. [41] have demonstrated that Precision, Recall, and F-measure are interval scales when you fix the length of the run N and the recall base RB; this is the case of a set of runs (assuming they retrieve the same number of documents) on a single topic (or on more topics, if they have the same recall base). In this case, Precision, Recall, and F-measure are a homomorphism with respect to the same ordering of runs in the ERS. However, if we mix together runs with different bounded lengths and/or different bounded recall bases, Precision, Recall and F-measure are no more interval scales, they are no more an affine transformation of each other and they even order the runs in different ways. Clearly, this is a severe issue when you need to average (or compute any other aggregate) across different topics or runs with different lengths.

Let us consider the universe set $S[N, K]$ which contains all the runs of any possible length n, less than or equal to N, and with respect to all the possible recall bases RB, less than or equal to K. To avoid trivial cases, we consider always N and K to be greater than 0. A run in $S[N, K]$ is represented by a triple $[r, n, RB]$, where r indicates the number of relevant documents retrieved by the run, n is the length of the run, and RB is the recall base, i.e. the total number of relevant documents for a topic. Note that, for each run in $S[N, K]$ it holds $n \leq N$ and $RB \leq K$ by construction, but we also have $r \leq (n \wedge RB)$, where $x \wedge y = \min\{x, y\}$, i.e. there is a (implicit) dependence on the recall base when it comes to the number of relevant retrieved documents.

We define $S_{n,RB}$ as the set which contains all the runs with the same length n and with respect to the same recall base RB. Therefore, we can express the universe set $S[N, K]$ as the union of such sets, namely

$$S[N, K] = \bigcup_{\substack{1 \leq n \leq N \\ 1 \leq RB \leq K}} S_{n,RB}$$

$S_{n,RB}$ models the typical case of runs all with the same length for a given topic (or for a set of topics which have the same recall base). This is exactly the case for which Ferrante et al. [41] have demonstrated that Precision, Recall and F-Measure are interval scales and an affine transformation of each other. However, this holds for each $S_{n,RB}$ separately while the issue we discuss in this section is what happens when you mix different $S_{n,RB}$, i.e. when you go towards $S[N, K]$.

1) PRECISION

Precision is equal to the fraction of the retrieved documents that are relevant. Therefore, for a run represented by triple $[r, n, RB]$, Precision is given by

$$Prec[r, n, RB] = \frac{r}{n}$$

Let us start from $S_{n,RB}$: $Prec$ maps this set into the set $\{0, 1/n, 2/n, \ldots, (n \wedge RB)/n\}$ and it has been proven by

Ferrante et al. that *Prec* is an interval scale in this case. However, already in this simpler case, there is a (implicit) dependency on the recall base, when it comes to the possible values of Precision. Therefore, even when we consider runs with the same length but for topics with different recall bases, i.e. S_{n,RB_1}, S_{n,RB_2}, S_{n,RB_3}, ...we are dealing with different scales, all embedded in the single interval scale whose image is $\{0, 1/n, 2/n, \ldots, (n \wedge \max\{RB_i\})/n\}$.

To understand the problems arising from mixing different lengths and recall bases, let us consider the general scenario of Precision defined on $S[N, K]$. This is the case where we consider the Precision measure defined on the set of the runs of any possible bounded length and recall base and we find that it is an interval scale only in the almost trivial cases of $N \leq 2$. Indeed, *Prec* maps $S[1, K]$, for any $K \geq 1$, into the set $\{0, 1\}$ and it is an interval scale since these values are equispaced. When $N = 2$, *Prec* maps $S[2, K]$, for any $K \geq 1$, into the set $\{0, 1/2, 1\}$; since the values are equispaced, Prec is still an interval scale. To compare the order induced on these sets by *Prec* (and the other measures), let us consider in more detail $S[2, 2]$. This set is

$$\{[0, 1, 1], [0, 1, 2], [0, 2, 1], [0, 2, 2], [1, 2, 1], [1, 2, 2], [1, 1, 1], [1, 1, 2], [2, 2, 2]\}$$

and $Prec[1, 2, 1] = Prec[1, 2, 2] = 1/2$, while $Prec[1, 1, 2] = 1$.

Continuing with a similar construction for $N = 3$, we obtain that *Prec* assumes the four possible values $\{0, 1/3, 1/2, 1\}$, when $K = 1$, and the five possible values $\{0, 1/3, 1/2, 2/3, 1\}$ when $K \geq 2$. Indeed, for runs of length at most 3, these are all the possible values of the fraction $\frac{r}{n}$ for $1 \leq n \leq 3$ and $0 \leq r \leq \min\{3, K\}$. Since these values are not equispaced, it is sufficient to state that *Prec* is not an interval scale on $S[3, K]$.

To prove that *Prec* is not interval on $S[N, K]$ for any finite $N > 3$, let us prove again that the values on the image are not equispaced. The three smallest values of $Prec[S[N, K]]$ are 0, $1/N$ and $1/(N-1)$. Indeed, the only other possible candidate to be the third smallest value, when $K \geq 2$, would be $2/N$, but $1/(N-1) < 2/N$ when $N > 2$. These three values are not equispaced since $1/N - 0 = 1/N \neq 1/N(N-1) = 1/(N-1) - 1/N$, when $N > 2$, and therefore *Prec* on $S[N, K]$ cannot be an interval scale when $N \geq 3$.

2) RECALL

The Recall measure depends explicitly on the recall base *RB*, i.e. the total number of relevant documents available for a given topic

$$Recall[r, n, RB] = \frac{r}{RB}$$

Note that for any admissible run $[r, n, RB]$, i.e. for which $r \leq n \wedge RB$, its recall value (implicitly) depends on n, creating a specular situation with respect to the one of Precision.

Recall is an interval scale on $S_{n,RB}$, since it is an affine transformation of *Prec*, as demonstrated by Ferrante et al.. However, due to the (implicit) dependency on n, even when we consider topics with the same recall base but runs with different length, i.e. $S_{n_1,RB}$, $S_{n_2,RB}$, $S_{n_3,RB}$, ...we are dealing with different scales, as discussed below.

Recall is an interval scale on the sets $S[N, 1]$ for any maximum length $N \geq 1$, since the image is the equispaced set $\{0, 1\}$. Applied to the sets $S[N, 2]$, for any N, *Recall* takes the values $\{0, 1/2, 1\}$ and, therefore, it is an interval scale as well. However, Precision and Recall induce, for example on $S[2, 2]$, two different orderings of the runs and so they stop to be an affine transformation of each other, i.e. they become two different interval scales. Indeed, consider the runs $[1, 2, 1]$ and $[1, 1, 2]$: we have seen that $Prec[1, 2, 1] = 1/2 < 1 = Prec[1, 1, 2]$, while it holds that $Recall[1, 2, 1] = 1 > 1/2 = Recall[1, 1, 2]$.

When we define *Recall* on $S[N, K]$, for $K > 2$, this measure is no more interval thanks to an argument similar to that used for Precision. Indeed, the two smallest non zero values of Recall on $S[N, K]$, are as $1/K$ and $1/(K-1)$, here obtained for a run with a unique relevant document with respect to a topic with $RB = K$ and $RB = K-1$, respectively.

Furthermore, it is immediate to see that Recall and Precision induce for any $N \geq 2$ and $K \geq 2$ two different orderings on $S[N, K]$, i.e. they become two different scales. Indeed, for any two runs $[r_1, n_1, RB_1]$ and $[r_2, n_2, RB_2]$, we have that $Prec[r_1, n_1, RB_1] < Prec[r_2, n_2, RB_2]$ if and only if $r_1/n_1 < r_2/n_2$, while $Recall[r_1, n_1, RB_1] > Recall[r_2, n_2, RB_2]$ if and only if $r_1/RB_1 > r_2/RB_2$. Both these condition are satisfied when

$$n_2/n_1 < r_2/r_1 < RB_2/RB_1$$

For example, if we take $r_1 = r_2$, $n_1 = 2\, n_2$ and $RB_2 = 2\, RB_1$, the previous condition is satisfied and the two runs are ordered in a different way by the two measures.

3) F_1 MEASURE

The F_1 measure is the harmonic mean of precision and recall

$$F_1[r, n, RB] = \frac{2}{(Prec[r, n, RB])^{-1} + (Recall[r, n, RB])^{-1}}$$

Some small algebra gives us that F_1 is also equal to

$$F_1[r, n, RB] = \frac{2\, r}{n + RB}$$

As before (see Ferrante et al.), we have that on $S_{n,RB}$, F_1 is an interval scale, being an affine transformation of *Prec* and *Recall*.

On the contrary, if we consider F_1 defined on $S[N, K]$, it is no more an interval scale, except for the almost trivial case $S[1, 1]$, whose image is the equispaced set $\{0, 1\}$. Let us first consider $S[2, 1]$ and $S[1, 2]$: in both these cases the values in the image of F_1 are no more equispaced, since F_1 takes the values $\{0, 2/3, 1\}$. If we consider F_1 on $S[2, 2]$, it takes the vales $\{0, 1/2, 2/3, 1\}$ and is still not an interval scale. Moreover, F_1 induces yet another ordering on $S[2, 2]$, since $F_1[1, 2, 1] = 2/3 = F_1[1, 1, 2]$, while it

holds that $Prec[1, 2, 1] = 1/2 < 1 = Prec[1, 1, 2]$ and $Recall[1, 2, 1] = 1 > 1/2 = Recall[1, 1, 2]$.

When we define F_1 on $S[N, K]$, for $N \geq 3$ or $K \geq 3$, this measure is no more an interval scale, as can be easily seen since the three smallest values of the image are 0, $2/(N + K)$ and $2/(N + K - 1)$, which are not equispaced. At the same time, using an example similar to the one used for *Recall*, we obtain that the ordering induced on $S[N, K]$ by F_1 in these latter cases differs from both the orderings induced by *Prec* and *Recall*.

4) SUMMARY AND DISCUSSION

We have demonstrated that, when regarding runs with a fixed length n and with respect to a fixed recall base RB, i.e. we consider $S_{n,RB}$ and runs of the same length for the same topic (or, more generally, topics with the same recall base), Precision, Recall, and F-measure are interval scales and they are an affine transformation of each other. As a consequence, they order runs in the same way and their Kendall's τ is 1.

However, when we start mixing runs with different length and/or with respect to different recall bases, the situation quickly gets more complicated. Only in the trivial (and not very useful in practice) case $N = 2$ and $K = 2$, i.e. $S[2, 2]$ where we have runs of length 1 or 2 and topics with 1 or 2 relevant documents, Precision and Recall are still both interval scales but they stop to be an affine transformation of each other. As a consequence, they order runs in different ways and their Kendall's τ is less than 1. F-measure already stops to be an interval scale and orders runs in yet another way than Precision and Recall, leading to a Kendall's τ less than 1. For $N > 2$ and $K > 2$ all of them (Precision, Recall, and F-measure) stop to be interval scales, departing from the interval assumption more and more, and they order runs in three completely different ways, again leading to a Kendall's τ less than 1. In the special case where we fix the length, Precision is still an interval scale, while if we fix a single recall base, Recall is still an interval scale, but in both cases the other measure is no more interval-scaled and also orders the runs in a different way.

We may be tempted to consider as positive the fact that sooner than later Precision, Recall, and F-measure start ordering runs in a different way and that their Kendall's τ is less than 1. Indeed, this is what we expect from evaluation measures, to embed different user models and to reflect different user preferences in ordering runs. This is also one of the main motivations why there is a debate and we would accept to derogate from requiring them to be interval scales: reflecting user preferences could be more important than complying with rigid assumptions.

However, we should carefully consider how this is happening. They initially are the ''same'' scale (except for an affine transformation), when we use them to measure objects with some shared characteristics, i.e. same run length n, and with respect to a similar context, i.e. same recall base RB. However, as soon as we measure objects with more mixed characteristics and contexts, they cease to be the ''same'' interval scale and only at that point they begin to order runs differently. This is more or less like saying that kilograms and pounds are the ''same'' interval scale only when we weigh people with the same height and from the same city but, as soon as we weigh people with different heights and/or coming from flatland or mountains, they become two different scales and they also possibly stop to be interval scales. This would sound odd and quite different from saying that weight and temperature are different (interval) scales because they measure different attributes/properties of an object or, in our terms, they would reflect different user preferences.

Why does this happen? Because run length n and recall base RB change. This is very clear and somehow more extreme in the case of the F-measure, where both n and RB explicitly appear in the equation of the measure.

We hypothesize that this could be even more severe and extreme in the case of rank-based measures since not only they combine, implicitly or explicitly, the two factors n and RB but they also mix them with the rank of a document and various discounting and accumulation mechanisms. Figure 3 gives a taste of this much more complex situation: it shows the simple (and somehow safe) case of $S_{4,4}$ and it already emerges how different are the behaviours and patterns in violating or complying with the interval scale assumption.

Why does this matter? As already said, because we need to aggregate scores across topics and runs and to compute significance tests. We do not only have the problem of how much evaluation measures violate the interval scale assumptions, required to compute aggregates, but also the issue of not mixing apples and oranges, i.e. scores from different scales, required to make aggregates sensible. In this respect, run length is a less severe issue which can be easily mitigated in practice, either by forcing a given length or because we are interested in lower cut-offs, e.g. 5, 10, 20, 30. The effect of the recall base can be mitigated by adopting measures that do not explicitly depend on it, even if the implicit dependency due to the capping of the image values will remain.

V. RELATED WORK

van Rijsbergen [110] was the first to tackle the issue of the foundations of measurement for IR by exploiting the representational theory of measurement. He observed that [110, pp. 365–366]

> The problems of measurement in information retrieval differ from those encountered in the physical sciences in one important respect. In the physical sciences there is usually an empirical ordering of the quantities we wish to measure. For example, we can establish empirically by means of a scale in which masses are equal, and which are greater or lesser than others. Such a situation docs not hold in information retrieval. In the case of the measurement of effectiveness by precision and recall, there is no *absolute* sense in which one can say that one particular pair of precision/recall values is better or

> worse than some other pair, or, for that matter that they are comparable at all

Later on, van Rijsbergen [112, p. 33] further stressed this issue: ''There is no empirical ordering of retrieval effectiveness and therefore any measure of retrieval effectiveness will be by necessity artificial''.

van Rijsbergen addressed this issue by exploiting the *additive conjoint measurement* [65], [70]. Additive conjoint measurement was a new part of the measurement theory developed as a reaction to the views of Campbell [22], [23] and the conclusions of the Ferguson Committee of British Association for the Advancement of Science [38], where Campbell was an influential member, which considered the *additive property*, i.e. the concatenation operation mentioned in Section III-A, as fundamental to science and proper measurement; as a consequence, measurement of psychological attributes, which is lacking such additive property, was not possible in a proper scientific way. As explained by Michel [76, p. 67]

> Conjoint measurement involves a situation in which two variables (A and B) are noninteractively [e.g. non additively] related to a third (X). It is not required that any of the variables be already quantified, although it is necessary that the values of X be orderable, and that values of A and B be independently identifiable (at least at a classificatory level). Then, via the order on P, ordinal and additive relations on A, B, and X may be derived

Typical examples from physics are the momentum of an object, which is affected by its mass and velocity, or the density, which is affected by its mass and volume [65].

van Rijsbergen considered retrieval effectiveness as the ''orderable X'' mentioned above and took precision P and recall R as the two variables A and B. In particular, he demonstrated that on the relational structure $(R \times P, \succsim)$ it was possible to define an additive conjoint measurement and to derive actual measures of retrieval effectiveness from it. Note that, in this way, he avoided the need to explicitly define what an ordering by retrieval effectiveness is and he considered that precision and recall contribute independently to retrieval effectiveness. The problem of how to order runs in the ERS has been addressed some years later by Ferrante et al. [39]–[41]. More subtly, van Rijsbergen treats precision and recall as two attributes which can be jointly exploited to order retrieval effectiveness but, each of them, is already a measure and quantification of retrieval effectiveness and, thus, this brings some circularity in the reasoning. Finally, van Rijsbergen did not address the problem of which are the scale properties of precision and recall (or other evaluation measures), which has been later addressed by Ferrante et al..

Bollmann and Cherniavsky [14], [15] built on the conjoint measurement work by van Rijsbergen and applied it to further study under which conditions the *MZ-metric* [57]. In particular, Bollmann and Cherniavsky leveraged what they called *transformational viewpoints*, i.e. elementary transformations of the runs which closely resemble the idea of swap and replacement used by Ferrante et al. much later on.

Bollmann [13] studied set-based measures by showing that measures complying with a monotonicity and an Archimedean axiom are a linear combination of the number of relevant retrieved documents and the number of not relevant not retrieved documents and how this could be related to collections and sub-collections. He thus addressed a problem somehow different from the one of the present paper, still leveraging the representational theory of measurement.

Busin and Mizzaro [19], Maddalena and Mizzaro [73] and Amigó and Mizzaro [5] proposed a unifying framework for ranking, classification, and clustering measures, which is rooted in the representational theory of measurement as well. They considered scales but as a way of mapping between relevance judgements (assessor scales) and Retrieval Status Value (RSV) (system scales) and of introducing axioms over them rather than a way of studying which are the scales actually used by IR evaluation measures and their impact on actual experiments.

As already discussed, Ferrante et al. [39] relied on the representational theory of measurement to formally study when evaluation measures are on an ordinal scale while Ferrante et al. [40], [41] proposed a more general theory of evaluation measures, proving when they are on an interval scale or not. Finally, Ferrante et al. [42] conducted a preliminary experimental investigation of the effects of IR measures being interval scales or not.

Even if not specifically focused on scales and their relationship to IR evaluation measures, there is a bulk of research on studying which constraints define the core properties of evaluation measures: Amigó et al. [6]–[9] and Sebastiani [99] face this issue from a formal and theoretical point of view, applying it to various tasks such as ranking, filtering, diversity and quantification, while Moffat [77] adopts a more numerical approach.

As it emerges from the above literature review, to the best of our knowledge, no one has dealt yet with the problem of considering the meaningfulness of IR experimental results and of transforming IR evaluation measures into interval scales.

VI. TRANSFORMING IR MEASURES TO INTERVAL SCALES

Let $(REL, \preceq)$ be a totally ordered set of relevance degrees, with a minimum called the non-relevant relevance degree $\mathrm{nr} = \min(REL)$ and a maximum $\mathrm{rr} = \max(REL)$; in the case of binary relevance, we set $REL = \{0, 1\}$ without any loss of generality. Let N be the length of a run, i.e. the number of retrieved documents, we call judged run $\hat{r}_t \in REL^N$ the vector of relevance degrees associated to each retrieved document, denoting by $\hat{r}[j]$ the j-th element of the vector.

Any IR evaluation measure M naturally defines an order among system runs. Indeed, taken any two runs $\hat{r}, \hat{s} \in REL^N$, we order them as follows

$$\hat{r} \prec \hat{s} \;\Leftrightarrow\; M(\hat{r}) < M(\hat{s}) \,. \tag{1}$$

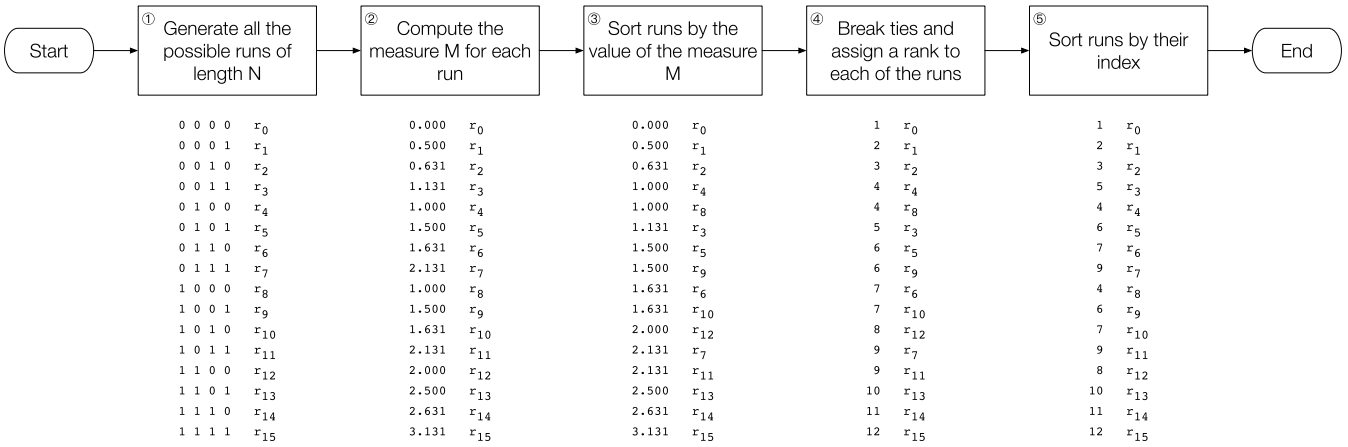


FIGURE 4. Procedure 1: flow chart for generating all the possible runs of length N and rank them according to a measure score.

Note that this is a *weak total order*, since $M(\hat{r}) = M(\hat{s})$ does not imply that $\hat{r} = \hat{s}$, and that it is the order called induced total order by Ferrante et al. Ferrante et al. [41]. It has the following characteristics, as discussed in the previous sections:

- it differs from measure to measure, i.e. each measure may produce a different RoS;
- it typically is not an interval scale, i.e. the produced values are not equi-spaced.

The basic idea of our approach is to keep the weak total order (1) produced by the measure M but making sure that all the possible values are equi-spaced.

The simplest way to achieve this result is to define first the nonlinear transformation φ from [0, 1] into $\mathbb{N}$ that maps each value m in the image of the measure M into its rank number. Then, we define the *ranked version of the measure*, i.e. the *interval-scaled version* of it, as $M_R = \varphi(M)$. Note that this approach is in line with what suggested by Gaito [50] to transform ordinal scales into interval ones.

Most of the measures are not one-to-one mappings and thus the cardinality of their image is strictly smaller than the cardinality of their domain, i.e. $|M(REL^N)| < |REL^N|$. The runs which are assigned the same value by the measure are called *ties*. As pointed out before, this is the reason why the order induced on REL^N by a measure in general is just a *weak* total order.

The map φ is then defined for any value m in the image $M(REL^N)$ as

$$\varphi(m) = |\{x \in M(REL^N) : x \leq m\}| \tag{2}$$

$M_R = \varphi(M)$ is an interval scale since the ranks are equi-spaced by construction; moreover, it preserves the RoS of M and thus it constitutes an interval-scaled version of it.

Finally, we have to deal with *tied* values in the measure. In statistics there are many ways of breaking ties [53]) and the most common are: average/mid, min, or max rank. However, each of these alternative strategies would result in a scale where the possible values are no longer equi-spaced. Indeed, suppose you have the following values $M = [0.00\,0.25\,0.40\,0.40\,0.70]$; the tied value 0.40 has ranks 3 and 4. If we chose the mid-rank tie breaking strategy, we would obtain $M_R = [1\,2\,3.5\,3.5, 5]$; using min-rank, we would obtain $M_R = [1\,2\,3\,3\,5]$; using max-rank, we would obtain $M_R = [1\,2\,4\,4\,5]$. In all these cases, the resulting scale would be no more equi-spaced.

Therefore, we simply eliminate the duplicate rank values and assign the same rank at all the tied positions, calling this tie-breaking strategy unique (`unq`). In the previous example, we would obtain $M_R[1\,2\,3\,3\,4]$.

For RBP with $p = 1/2$ we have that $\varphi(m) = 2^N m$ while for RR we have that $\varphi(m) = N + 1 - \frac{1}{m}$. However, in general, the function φ does not have any analytical expression, it is **nonlinear** and it varies from measure to measure.

A. IMPLEMENTATION

The approach described in the previous section is implemented in two procedures:

- generating all the possible runs of length N and ranking them according to a measure score, shown in the flow chart in Figure 4. This step computes the ranked version of the measure $M_R = \varphi(M)$ for each of all the possible runs of length N.
- for each topic of a run r determine its rank, shown in the flow chart in Figure 5. This steps computes the ranked version of the measure $M_R = \varphi(M)$ for each topic of a run r.

Let us consider the first procedure, shown flow chart in Figure 4; as an example, we assume run length $N = 4$, binary relevance judgements, and DCG with $\log_2$ as measure, corresponding to Figure 3g. The first step is to generate all the possible runs: since we are working with binary relevance judgments there are $2^N = 16$ possible runs, shown in the bottom part of Figure 4, and indexed from r_0 to r_{15}. Note that the index of each run is just the decimal encoding of the binary

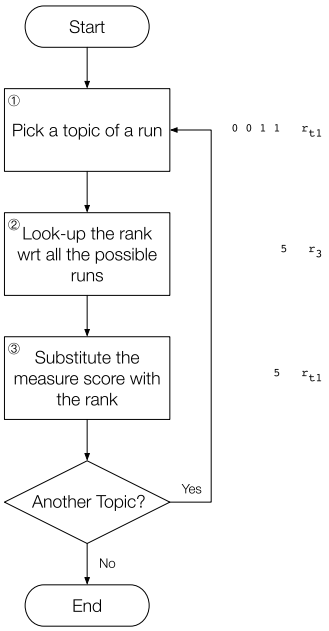


FIGURE 5. Procedure 2: flow chart for determining the rank for each topic of a run.

vector corresponding to the run; this will then be exploited in the next procedure. In the second step, we compute the measure M for each run; in our example, we compute DCG and we actually obtain 12 distinct values of DCG. In the third step, we sort[9] the runs by the value of the measure M. Then, in the fourth step, we break ties and we assign a rank to each run; in the example at the bottom of we used the unique strategy (`unq`) described above. Note that this step determines the ranked version of the measure $M_R = \varphi(M)$ for each of all the possible runs; in our example, we have that: $\varphi(3.131) = \varphi\big(M(1,1,1,1)\big) = 12$, $\varphi(2.631) = \varphi\big(M(1,1,1,0)\big) = 11$, $\varphi(2.500) = \varphi\big(M(1,1,0,1)\big) = 10$, $\varphi(2.131) = \varphi\big(M(1,0,1,1)\big) = \varphi\big(M(0,1,1,1)\big) = 9$ and so on until $\varphi(0.500) = \varphi\big(M(0,0,0,1)\big) = 2$, and $\varphi(0.000) = \varphi\big(M(0,0,0,0)\big) = 1$.

Let us consider the second procedure, shown flow chart in Figure 5; as an example, we assume a run consisting of two topics $r_{t_1} = (0,0,1,1)$ and $r_{t_2} = (1,0,0,1)$; we have that $M(r_{t_1}) = 1.131$ and $M(r_{t_2}) = 1.500$. In the first step, we pick a topic of the run, say t_1. In the second step we look up the rank of this topic with respect to all the possible runs, i.e. the fourth step of the previous procedure. To simplify this step, we exploit the fact that the decimal encoding of the binary vector of the run corresponds to its index; in our case $0011_2 = 3_{10}$, so r_{t_1} corresponds to r_3 in all the possible runs, whose rank is 5. In the third step, we substitute the measure score, i.e. $M(r_{t_1}) = 1.131$, with its rank, i.e. $M_R(r_{t_1}) = \varphi\big(M(r_{t_1})\big) = 5$; in other terms we determine the ranked version of the measure for that topic. These steps are repeated for all the topics constituting a run; in our case, for topic t_2 we obtain that r_{t_2} corresponds to r_9, whose rank is 6, and thus $M_R(r_{t_2}) = \varphi\big(M(r_{t_2})\big) = 6$.

[9]In our implementation we opted for ascending order but descending order would have worked the same.

B. LIMITATIONS: RUNS OF DIFFERENT LENGTH

When working with runs all with the same length, the proposed approach maps a measure into a proper interval scale, actually the same scale for all the runs, and this allows us to compute aggregates across runs with the same length.

If we work with runs of different length, the proposed approach maps each length (run) into a proper interval scale but it differs from length to length. For example, in the case of DCG with $\log_2$ there are 24 distinct values for $N = 5$, 768 for $N = 10$, 24576 for $N = 15$, and so on; all of them correspond to a ranked measure (interval scale) with a different number of steps. As a consequence, even using our approach, we cannot aggregate across runs with different length.

However, as already discussed, this is a problem easily manageable in practice. Indeed, for small run lengths or low cut-offs of typical interest, such as the top 10 documents, it is reasonable to assume that the runs have all the same length, since runs are usually able to retrieve enough documents. In the more general case and for bigger run lengths, if a run does not retrieve enough documents, it could be padded with not relevant documents. Therefore, we can consider our approach as generally applicable with respect to this issue.

C. LIMITATIONS: DIFFERENT TOPICS

Let us now assume that we have fixed a run length which is the same for all the runs and which allows us to compute aggregates across runs. What happens if we need to compute aggregates across topics?

1) MEASURES NOT DEPENDING ON THE RECALL BASE

In the case of measures not depending on the recall base, since the length of the run is the same for all the runs across all the topics, our approach maps a measure into the same interval scale for all the runs and all the topics. Therefore, we can safely compute aggregates also across topics.

2) MEASURES DEPENDING ON THE RECALL BASE

In the case of the measures depending on the recall base, as already explained, due to the recall base changing from topic to topic, there exists no single (interval) scale which can be used across all the topics. As a consequence our approach could not be directly applied. However, we could use it as a surrogate that brings, at least, some more "intervalness" to a measure.

Indeed, on each single topic, our approach maps a measure depending on the recall base into a proper interval scale, whose steps are equi-spaced. When we deal with two (or more) topics, we would need to find an interval scale where it is possible to match the steps from the (two) scales of each topic into some "bigger" set of equi-spaced steps, which can accommodate all of them. However, as shown in

Example 11 and in Section IV-D, this common super-set of steps does not exist, with the exception of the trivial cases mentioned.

Therefore, as an approximation, we can pretend that the scale for each topic is the overall common scale – and, as said before, this is exactly what happens in the case of measures not depending on the recall base – and use it across topics, even if this will actually stretch the steps of each topic.

Example 12 (Surrogating Recall to an Interval Scale): Suppose we are dealing with runs of length $N = 2$*, i.e.* $r_0 = [0\,0]$, $r_1 = [0\,1]$, $r_2 = [1\,0]$, $r_3 = [1\,1]$. *If the recall base for the first topic* q_1 *is* $RB_1 = 2$*, these runs are mapped to the Recall values* $\{0, \frac{1}{2}, \frac{1}{2}, 1\}$*; if the recall base for the second topic* q_2 *is* $RB_2 = 3$*, these runs are mapped to the Recall values* $\{0, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}\}$.

Our transformation approach maps the runs of both topics to $\{1, 2, 2, 3\}$*, which is a proper interval scale on each topic separately. However, if we look at the two topics together and we use this mapped scale, we are slightly stretching the steps of the original scales. For example, if we compute the difference between* r_2 *and* r_3*, on this mapped scale it is same, i.e.* 1*, for both* q_1 *and* q_2 *while on the original scales it is* $\frac{1}{2}$ *for* q_1 *and* $\frac{1}{3}$ *for* q_2*. This means that our transformation also effects ordinal scales when a recall base is involved.*

D. LIMITATIONS: IMPLEMENTATION

A practical limitation of the proposed methodology is the need to generate all the possible runs to determine their rank, as shown in the flow chart in Figure 4.

To give an idea of the computational resources required, runs of length $N = 30$ mean an occupation of $2^{30} * 30 * 8 =$ 240 GByte of memory, just for holding all the possible runs. A length $N = 40$ would mean 320 TByte of memory, which is not feasible in practice.

The code is implemented in Matlab and thus we considered 8 bytes for representing a digit, since this is the size of a double. Even if we considered a more compact representation, in some other language like C, using just 1 bit per digit, it would have meant 40 TByte of memory for runs of length $N = 40$.

The case of multi-graded relevance is left for future work, due to its exponential explosion in the number of possible cases. For example, switching from a binary to a 4-valued scale, for the run length $N = 30$ the number of possible runs would grow from $2^{30} \sim 10^9$ to $4^{30} = 2^{60} \sim 10^{18}$.

VII. EXPERIMENTAL SETUP

We regard the following evaluation measures: Precision (P) and Recall (R) [111], AP [17], RBP [78], DCG and nDCG [60] and RR [102]. We calculated RBP by setting $p \in \{0.3, 0.5, 0.8\}$, indicated respectively as RBP_p03, RBP_p05, and RBP_p08; for DCG and nDCG we use a log_2 and a log_{10} discounting, indicated respectively as DCG_b02, nDCG_b02, DCG_b10, and nDCG_b10.

We considered the following datasets:

- *Adhoc Track* `T08` *[120]:* 528,155 documents of the TIPSTER disks 4-5 corpus minus congressional record; `T08` provides 50 topics, each with binary relevance judgments and a pool depth of 100; 129 system runs retrieving 1,000 documents for each topic were submitted to `T08`.
- *Common Core Track* `T26` *[3]:* 1,855,658 documents of the New York Times corpus; `T26` provides 50 topics, each with multi-graded relevance judgments (not relevant, relevant, highly relevant); relevance judgements were done mixing depth-10 pools with multi-armed bandit approaches [69], [118]; 75 system runs retrieving 10,000 documents for each topic were submitted to `T26`.
- *Common Core Track* `T27` *[4]:* 595,037 documents of the Washington Post corpus; 50 topics, each with multi-graded relevance judgments (not relevant, relevant, highly relevant); relevance judgements were done adding stratified sampling [27] and move-to-front [31] approaches to the `T26` procedure; 72 system runs retrieving 10,000 documents for each topic were submitted to `T27`.

In the case of the `T26` and `T27` tracks we mapped their multi-graded relevance judgement to binary ones using a lenient strategy, i.e. whatever above not relevant is considered as relevant.

For each track we experimented with the run lengths $N \in \{5, 10, 20, 30\}$, i.e. we cut runs at the top-N retrieved documents. In terms of our transformation methodology, this means considering a space of possible runs containing, roughly, $\{32, 10^3, 10^6, 10^9\}$ runs, respectively. We indicated the run length in the identifier of the track, e.g. `T08_10` means `T08` runs cut down at length 10.

In significance tests, we used a confidence level $\alpha = 0.05$.

To ease the reproducibility of the experiments, all the source code needed to run them is available in the following repository: https://bitbucket.org/frrncl/ieeea2021-fff/src/master/.

VIII. EXPERIMENTS

In this section we compare what happens when using a proper interval scale to what happens when not using an interval scale; in the case of IR measures, when they are not interval scales, they are, at least, ordinal scales. We perform this comparison from three different points of view, detailed below and discussed in dedicated Sections. In essence, we use our *ranked version of the measure* M_R as interval scale, since it guarantees by construction to be an interval scale, and we contrast it to the IR evaluation measure M it is derived from. When we the original IR evaluation measure is already an interval scale, we observe the same behaviour for both M and M_R; when the original IR evaluation measure is not an interval scale, we observe different behaviours.

More in detail, in Section VIII-A we validate our approach and answer the research question "How far is a measure from being an interval scale?". Then, in the next two sections we

investigate the effects of using or not a proper interval scale in IR experiments. In particular, in Section VIII-B we study how this affects the correlation among evaluation measures, i.e. the main tool we use to determine how close are two measures. In Section VIII-C we analyse how this impacts on the significance tests, both parametric and non-parametric, i.e. the main tool we use to determine when IR systems are significantly different or not.

The overall objective of the experimentation is to provide a first quantitative assessment of the impact of departing (or not) from scale assumptions. As already anticipated, these experiments motivate the need for and call for a more careful consideration, in the IR field, of the scale assumptions and the limitations they may pose on our experimental findings. The goal of this section is to show neither that a specific evaluation measure is broken nor that previous findings in the literature are potentially wrong.

The following sections report, separately, the case of measures not depending on the recall base – namely, RBP, DCG, RR, and P – and measures depending on the recall base – namely, AP, nDCG, and R. Indeed, as previously explained, with a fixed run length, in the former case it is possible to find an overall interval scale, which is the same across all the topics, and apply our transformation approach in an exact way; in the latter case, an overall interval scale common to all the topics does not exist and our transformation is just the best surrogate that can be figured out.

A. CORRELATION BETWEEN MEASURES AND THEIR RANKED VERSION. HOW FAR IS A MEASURE FROM BEING AN INTERVAL SCALE?

In this section we study the relationship between each measure and its ranked version, i.e. its mapping towards an interval scale, as explained in Section VI. This analysis allows us: 1) to validate our approach, verifying that it produces the expected results; 2) to understand how much a measure changes when it is transformed, seeking an explanation for this change; 3) to understand what happens when we apply our transformation approach in a ''surrogate'' way in the case of measures depending on the recall base.

We compute both the overall and the topic-by-topic Kendall's τ correlation between each measure M and its ranked version.[10] Ferro [43] has shown that, even if the absolute correlation values are different, removing or not the lower quartile runs produces the same ranking of measures in terms of correlation; similarly, he has shown that both τ and AP correlation τ_{ap} [125] produce the same ranking of measures in terms of correlation. Therefore, we focus only on Kendall's τ without removing lower quartile systems.

As explained in Section IV-C, the *topic-by-topic correlation* is expected to be always 1.0, since the ranked version of a measure preserves the same order of runs on each topic by construction. As a sanity check, we verified that the topic-by-topic correlation is indeed 1.0 in all the cases and we do not report it in the following tables for space reasons. On the contrary, the *overall correlation*, i.e. the traditional one, can be different from 1.0 for the reasons discussed above: the preliminary average operation would be not allowed in the case of an original measure which is not an interval scale while it is allowed in the case of the corresponding ranked measure; also, different recall bases across topics can lead to different scales which should not be averaged together.

In general, we can consider the overall Kendall's τ correlation between a measure and its ranked version as a *proxy* providing us with an estimation of how far a measure is from both being a proper interval scale and, in the case of measures depending on the recall base, also being safely averaged across topics. Note that this approach is in line and extends what was proposed by Ferrante et al. [42] when they suggested to use the overall Kendall's τ correlation between a measure and the Set-Based Total Order (SBTO) and Rank-Based Total Order (RBTO) interval scales as an estimation of how much a measure is an interval scale.

Table 2 summarizes the outcomes of the overall correlation analysis between measures and their ranked version, e.g. we computed Kendall's τ correlation between DCG and DCG_R, the interval-scaled version of DCG according to the approach we described in Section VI.

From a very high level glance at Table 2 we can see that Kendall's τ correlation changes due to the transformation but not in a too excessive way which suggests that we are not running into any pathological situation.

1) MEASURES NOT DEPENDING ON THE RECALL BASE

As previously discussed, Precision is already an interval scale – a different scale for each run length but, fixed a length, the same scale for all the topics, allowing to safely average across them. In this case, our transformation is just a mapping between interval scales, as the transformation between Celsius and Fahrenheit is. As the overall Kendall's τ correlation is always 1.0, we are experimentally confirming the correctness of our transformation approach and that everything is working as expected.

The other case in which we see this happening is RBP_p05, which we already know to be an interval scale, but different from the one of Precision.

On the other extreme, there is RR, which is the farthest away from being an interval scale, among the measures not depending on the recall base. The overall Kendall's τ correlation is in the range $0.67 - 0.93$ which is systematically lower than the correlation of all the other measures in this group – namely P, RBP, and DCG. This suggests that transforming RR into an interval scale requires a more marked correction or, in other terms, that it experiences a drop in ''intervalness'' in the range $7\% - 33\%$. We can also observe another stable pattern in the case of RR: the bigger the length of the run, the lower the correlation between RR and its ranked version. This suggests that RR departs more and more from the

[10] To avoid errors due to floating point computations, we rounded averages to 8 decimal digits.

TABLE 2. Kendall's τ overall correlation analysis between each measure and its respective ranked version, using the `unq` tie breaking approach.

Track	P	RBP_p05	RR	RBP_p03	RBP_p08	DCG_b02	DCG_b10	R	AP	nDCG_b02	nDCG_b10
T08_05	1.0000	1.0000	0.9211	0.9522	0.9605	0.9759	1.0000	0.8145	0.8219	0.9759	1.0000
T08_10	1.0000	1.0000	0.8466	0.9500	0.9527	0.9553	1.0000	0.8030	0.8243	0.9541	0.9965
T08_20	1.0000	1.0000	0.7677	0.9498	0.9498	0.9334	0.9537	0.7943	0.8197	0.9285	0.9452
T08_30	1.0000	1.0000	0.7329	0.9498	0.9508	0.9128	0.9261	0.7948	0.8377	0.9072	0.9147
T26_05	1.0000	1.0000	0.9219	0.9706	0.9661	0.9717	1.0000	0.6974	0.6903	0.9717	1.0000
T26_10	1.0000	1.0000	0.8232	0.9704	0.9610	0.9567	1.0000	0.8207	0.7633	0.9582	0.9982
T26_20	1.0000	1.0000	0.7500	0.9704	0.9560	0.9517	0.9690	0.8848	0.8600	0.9560	0.9661
T26_30	1.0000	1.0000	0.6725	0.9704	0.9582	0.9264	0.9127	0.8901	0.8701	0.9264	0.9141
T27_05	1.0000	1.0000	0.9350	0.9597	0.9536	0.9730	1.0000	0.7540	0.8312	0.9730	1.0000
T27_10	1.0000	1.0000	0.8830	0.9601	0.9476	0.9436	1.0000	0.7860	0.8442	0.9491	0.9912
T27_20	1.0000	1.0000	0.8227	0.9601	0.9288	0.9272	0.9358	0.7929	0.8309	0.9155	0.9068
T27_30	1.0000	1.0000	0.7958	0.9601	0.9303	0.9295	0.9397	0.8191	0.8380	0.9068	0.9139

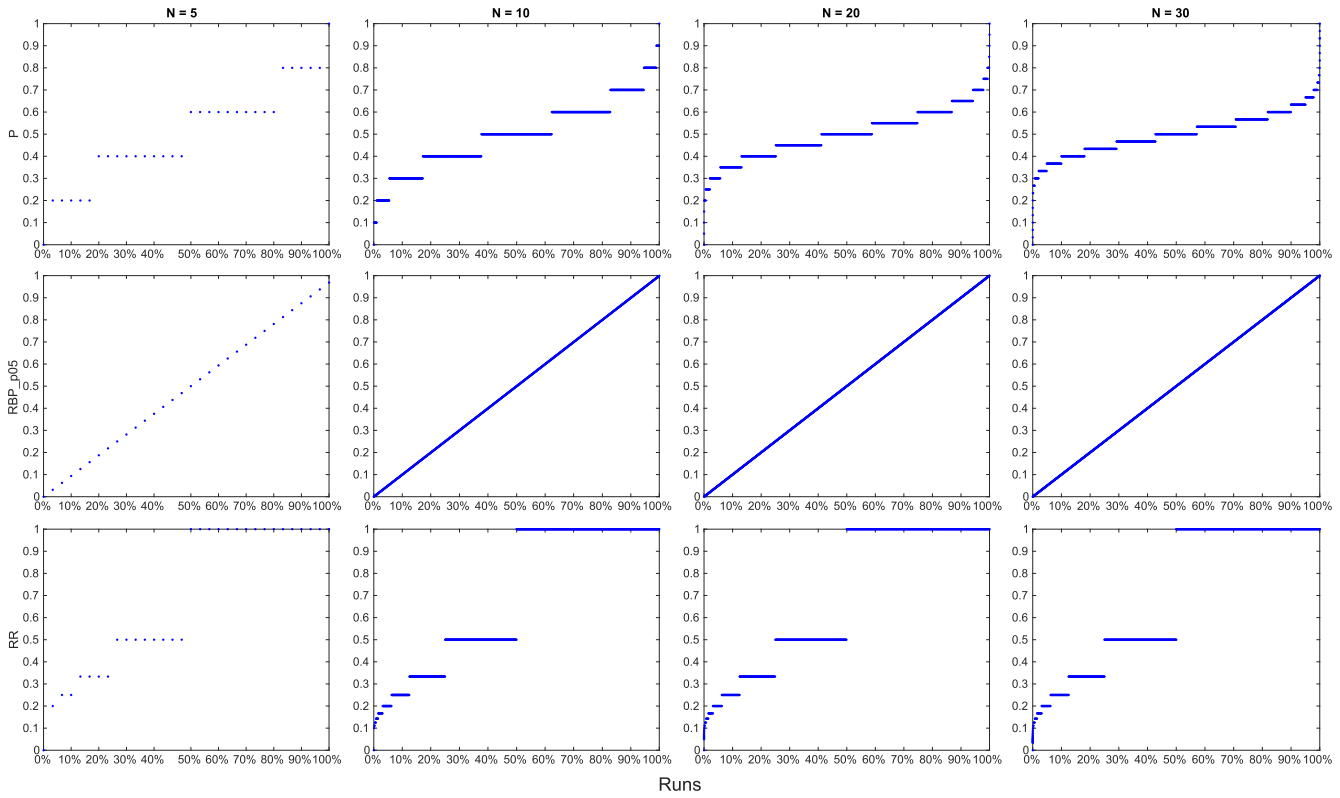


FIGURE 6. All the possible values of P, RBP with $p = 0.5$, and RR for runs of length {5, 10, 20, 30}.

interval scale assumption as the run length increases; we will now see why this is happening.

Figure 6 plots the values of P, RBP with $p = 0.5$, and RR for all the possible runs of a given length. On the X axis there are the runs, increasingly ordered by the value of the measure; this is the order of runs considered by the ranked version of the measure which then just equi-spaces the values and removes ties. The Y axis reports the value of the measure for each run. The labels on the X axis report the fraction of runs up to that point; so, for example, in the case of P and $N = 5$, we can understand that 20% of the runs receive the value $P = 0.2$, 30% the value $P = 0.4$, 30% the value $P = 0.6$, and 20% the value $P = 0.8$; 1 run gets the value 0.0 and 1 run the value 1.0.

RBP with $p = 0.5$ yields distinct equi-spaced values for each of the possible runs or, in other terms, it produces equi-spaced clusters of values containing one single value in each cluster. In the case of P, we can see how the clusters are still equi-spaced but they contain tied values, visible as horizontal segments. Finally, in the case of RR, not only the clusters are not equi-spaced – and this breaks the interval scale assumption – but they also increase more and more, and only in one region of the range, as the run length increases, making RR less and less interval. Indeed, the number of clusters is not equi-spaced but always constant to 5 in the range [0.2, 1.0], independently from the run length; on the other hand, in the range [0, 0.2) it increases from 1 to 6, 16, and 26 as the run length increases.

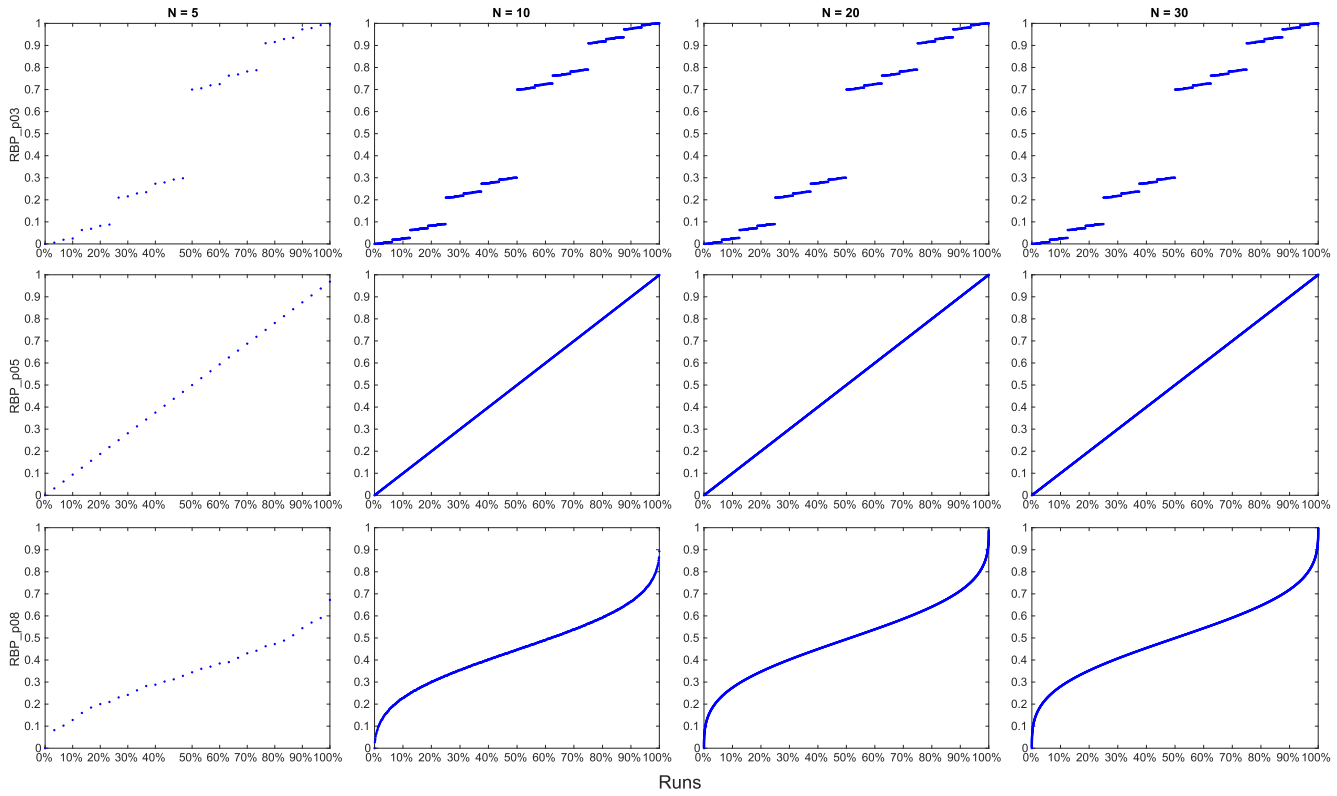


FIGURE 7. All the possible values of RBP with $p \in \{0.3, 0.5, 0.8\}$ for runs of length $\{5, 10, 20, 30\}$.

Moreover, Figure 6 visually shows us why different run lengths correspond to different scales – interval or not depending on whether clusters are equi-spaced or not. In all the cases, the number of clusters increases as the length of the run increases and this makes the scales to be different. Note that this behaviour is not like getting a more and more accurate scale, which would be a desirable property, but it is rather like saying that the height scale would become denser and denser as you measure taller and taller people. Therefore, as already said, we should avoid to mix values of measures coming from runs with different lengths, e.g. by averaging.

When it comes to RBP, we know from Ferrante et al. [40], [41] that: RBP_p05 is an interval scale; RBP_p03 is an ordinal scale keeping the same ordering as RBP_p05 but being no more an interval scale; and, RBP_p08 uses a different ordering from RBP_p05 and it is not an interval scale. This is also reflected in the overall correlation values. As already noted, and expected, the overall correlation for RBP_p05 is always 1.0 while it drops in the range $0.95 - 0.97$ for RBP_p03. RBP_p03 and RBP_p05 order runs in the same way, which also means that their ranked version is the same. Therefore, the $3\% - 5\%$ difference between RBP_p03 and RBP_p05 depends only on the lack of equi-spacing of RBP_p03 and the problems it causes when averaging. This also means that this drop in "intervalness" of RBP_p03 is not the effect of a user model somehow different from the one of RBP_p05, possibly resulting in a different order of the systems, which is the typical explanation provided in these cases instead. In the case of RBP_p08 we observe a similar behaviour and the correlation drops in the range $0.93 - 0.96$ with an "intervalness" loss in the range $4\% - 7\%$.

When it comes to increasing run lengths, we can observe that the correlation values of RBP oscillate a bit, they tend to get more stable as the run length increases and this happens more for RBP_p08 than for RBP_p03. While this still might be partially due to the measure being more or less interval scale depending on the run length, we think that in the case of RBP this is mostly motivated by another reason. Indeed, as previously discussed, RBP does not use the full range [0, 1] because of the $\frac{1}{1-p}$ overestimation which impacts more as p increases and the length of the run is smaller. Therefore, we think that the increase in range of RBP is the motivation of the observed small changes in the correlation values. We can clearly see this behaviour in Figure 7 for RBP_p08 whose values fall in the full range [0, 1] only for $N = 20$ and $N = 30$, while this effect is mostly negligible for RBP_p03 and RBP_p05. As a consequence, correlation values tend to get more stable for $N = 20$ and $N = 30$ in the case of RBP_p08 while they are quite stable for RBP_p03 and RBP_p05, independently from the run length.

Finally, when it comes to DCG, we can observe from Table 2 that its overall correlation is above 0.9 with an "intervalness" loss in the range $2\% - 9\%$, suggesting it is moderately departing from its ranked version. We can also observe for DCG_b10 that the correlation for run lengths $N = 5$ and $N = 10$ is always 1.0, which may look surprising;

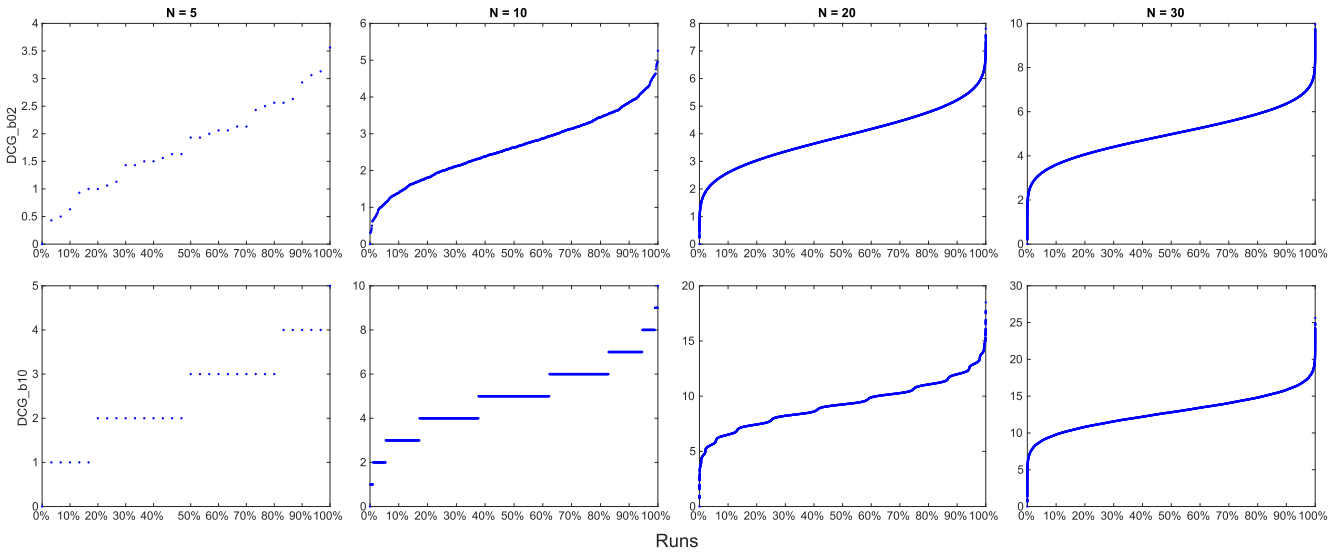


FIGURE 8. All the possible values of DCG with log base $b = 2$ and $b = 10$ for runs of length $\{5, 10, 20, 30\}$.

this is actually an artefact of the log base 10 which causes the discount to be applied from the 11th rank onwards. Therefore, for run lengths up to the log base, DCG is basically counting the number of relevant retrieved documents, as it is clear from Figure 8, and this produces the same interval scale as P. However, we should be aware of this somehow unusual behaviour of DCG_b10 because it changes from being an interval scale for runs up to length 10 to not being it anymore afterwards.

Moreover, as a general trend, DCG tends to be less and less an interval scale as the run length increases. If we look at the possible values of DCG in Figure 8, this may sound surprising since DCG visually behaves in a very similar way to RBP_p08, at least after the run length is big enough to compensate for possible effects of the log base itself. However, while RBP_08 does not have tied values, DCG exhibits an increasing number of tied values, clustered unevenly across the range – this is not visible from the figure, especially for DCB_b02, due to the small size of the tied clusters but we have verified it on the numerical data underlying the plot. Therefore, the increasing number of uneven tied cluster explains, as in the case of RR, why DCG is less and less an interval scale.

2) MEASURES DEPENDING ON THE RECALL BASE

Let us go back to Table 2 and consider the case of R. We know that Precision and Recall, on each topic separately, are already interval scales and just a transformation of the same interval scale. Therefore, when we map them to their ranked version, it is actually the same interval scale for both of them and it is yet another transformation of their common original interval scale. However, while this means an overall correlation 1.0 in the case of Precision, it drops in the range $0.7-0.9$ in the case of Recall. This $10\% - 30\%$ loss in "intervalness" is entirely due to the effect of the recall base and let us understand how careful we should be before averaging across topics.

AP follows a somehow similar pattern with overall correlation values in the range $0.69 - 0.86$ with an "intervalness" loss in the range $14\% - 31\%$.

On the other hand, nDCG exhibits overall correlation values very close to those of DCG, all above 0.9 with an "intervalness" loss in the range $2\% - 10\%$. We observe another somehow surprising behaviour of nDCG_b10: for runs of length $N = 5$ the correlation is always 1.0, indicating that it is an interval scale and, most of all, that there is no effect of the recall base. The fact is that on all the tracks under examination, all the topics have at least 5 relevant documents, so the recall base is never below 5; when you trim runs to length $N = 5$, the DCG_b10 of the ideal run, i.e. the factor used to normalize DCG in nDCG, is constant to 5 for all the topics and so there is no recall base effect for this reason. On the other hand, there is 1 topic with less than 10 relevant documents on both `T08` and `T26` and 4 topics on `T27`. As a consequence, nDCG_b10 drops slightly below 1.0 on `T08_10` and `T26_10` and a bit more on `T27_10`. This further stresses the need to be careful, or at least aware of, the fact that DCG/nDCG may change behaviour and nature for document cut-offs below the log base. Moreover, this gives us an idea of how much even very small changes in the recall base can have an impact and how careful we should be when aggregating across topics.

Finally, both DCG and nDCG are mapped to the same ranked measure, exactly as P and R are mapped to the same ranked measure. However the loss of "intervalness" of R is much bigger than the one of nDCG. We hypothesise that this is due to how the recall base is accounted for in the measure: in the case of R it is a straight division by the recall base itself while in the case of nDCG it is a division by the DCG of the ideal run, which is also one of the possible runs considered in the mapping. The latter is a much smoother normalisation than just an integer number representing the total number of relevant documents. The behaviour of AP, very close to

TABLE 3. Kendall's τ overall correlation analysis between each measure and its respective ranked version, using the mid-rank tie breaking approach.

Track	P	RBP_p05	RR	RBP_p03	RBP_p08	DCG_b02	DCG_b10	R	AP	nDCG_b02	nDCG_b10
T08_05	0.9680	1.0000	0.9694	0.9522	0.9605	0.9622	0.9680	0.7973	0.8208	0.9622	0.9680
T08_10	0.9392	1.0000	0.9713	0.9500	0.9527	0.9462	0.9392	0.7725	0.8213	0.9450	0.9362
T08_20	0.9043	1.0000	0.9736	0.9498	0.9498	0.9276	0.9045	0.7809	0.8161	0.9227	0.8970
T08_30	0.8978	1.0000	0.9743	0.9498	0.9508	0.9094	0.8931	0.7625	0.8401	0.9038	0.8851
T26_05	0.9670	1.0000	0.9794	0.9706	0.9661	0.9614	0.9670	0.6816	0.6907	0.9614	0.9670
T26_10	0.9624	1.0000	0.9782	0.9704	0.9610	0.9502	0.9624	0.8153	0.7540	0.9502	0.9599
T26_20	0.9245	1.0000	0.9854	0.9704	0.9560	0.9488	0.9242	0.8615	0.8557	0.9517	0.9257
T26_30	0.8785	1.0000	0.9832	0.9704	0.9582	0.9228	0.8831	0.8319	0.8665	0.9228	0.8846
T27_05	0.9535	1.0000	0.9782	0.9597	0.9536	0.9607	0.9535	0.7501	0.8294	0.9607	0.9535
T27_10	0.9432	1.0000	0.9827	0.9601	0.9476	0.9272	0.9432	0.7628	0.8521	0.9327	0.9354
T27_20	0.9086	1.0000	0.9843	0.9601	0.9288	0.9225	0.9084	0.7352	0.8356	0.9076	0.8873
T27_30	0.9103	1.0000	0.9835	0.9601	0.9303	0.9264	0.9100	0.7650	0.8395	0.9037	0.8873

the one of R, supports this intuition, since also AP adopts a straight division by the recall base itself.

3) IMPACT OF THE TIE BREAKING STRATEGY

In this section, we perform a further validation of our mapping approach. As explained in Section VI, evaluation measures often produce tied values and we remove these tied values by assigning them their unique rank position, since this ensures that values are kept equi-spaced. However, as pointed out by Gibbons and Chakraborti [53], there are many other common ways of breaking ties, one of which if the mid-rank strategy, i.e. keeping the average of the ranks of the tied values.

Table 3 shows what happens to our transformation approach when using the mid-rank tie breaking instead of the unq one used in Table 2. Let us consider Precision whose overall correlation values drops from 1.0 to the range $0.90 - 0.97$. Since Precision is already an interval scale, this drop is entirely due to the fact that the mid-rank tie breaking strategy produces a scale whose values are not equi-spaced anymore and thus a scale which is not interval anymore. Only RBP_p05 keeps the overall correlation 1.0 because it is already an interval scale but it does not have any tied value, so it is insensitive to the tie breaking strategy. As a general trend, we can see that the overall correlation values in Table 3 are lower than those in Table 2 due to the loss of "intervalness" caused by the tie breaking strategy.

Therefore, we validated that the appropriate way of implementing our transformation approach is by using the unq tie breaking strategy. Moreover, this further stresses how much the lack of equi-spaced values, lack due to any reason, impacts on our measurement process.

B. CORRELATION AMONG MEASURES AND AMONG THEIR RANKED VERSIONS. UNVEILING THE "TRUE" CORRELATION AMONG EVALUATION MEASURES

Table 4 summarizes the outcomes of the correlation analysis among measures and among their ranked versions, i.e. on the one side we compute Kendall's τ overall correlation among all pairs of measures, on the other side we compute Kendall's τ overall correlation among the same pairs of ranked measures. In this way, we can study whether and how the estimated relationship among measures changes when passing to their ranked version or, in other terms, to what extent being an interval scale or not biases our estimations. In particular, the column $\Delta\%$ reports the percent increase/decrease of the correlation between the ranked measures (labelled RnkMsr) with respect to the correlation between the original measures (labelled Msr), i.e. how much the correlation between two measures is underestimated/overestimated due to the fact that a measure is not an interval scale. Table 4 reports results for the T08_30, T26_30, and T27_30 tracks; results for the other tracks are similar but not shown here for space reasons.

We can observe from Table 4, as very coarse and general trends, that correlation is overestimated ($\Delta\%$ column negative) in the range $[-33.30\%, -0.16\%]$, i.e. two evaluation measures are less close to each other than what we would be induced to think; conversely, correlation is underestimated ($\Delta\%$ column positive) in the range $[0.4\%, 25.82\%]$, i.e. two evaluation measures are closer to each other than what we would be induced to think. This observation opens up a relevant question for IR experiments: are IR measures really that different? Do we need all of them? Are we really scoring runs according to different user viewpoints or are these differences just an artefact of violating the scale assumptions? How much of what reported in the literature is due just to this scale violation bias?

In the following sections we discuss a few examples from Table 4 of how the correlation may change.

1) MEASURES NOT DEPENDING ON THE RECALL BASE

Let us start from the correlation between Precision and RBP with $p = 0.5$. We already know that they are interval scales and, therefore, their ranked version is just another mapping of their respective interval scales – and this is why in Table 2 their overall correlation is 1.0. We can observe from Table 4 that on T08_30 the correlation between RBP_p05 and P is 0.7858 and, as expected, the correlation between their ranked versions is the same, since the interval scale behind the original measures and their ranked version is the same. The same happens for the other tracks, i.e. T26_30 and T27_30.

Note that "the correlation between Precision and RBP with $p = 0.5$ on T08_30 is 0.7858" is an example of a *meaningful statement* in IR, since it is invariant to a permissible transformation of the interval scales of these two measures and it does not change its truth value.

TABLE 4. Kendall's τ overall correlation analysis between each pair of measures (labelled `Msr`) and between each pair of ranked measures (labelled `RnkMsr`) on tracks `T08_30`, `T26_30`, and `T27_30`, using the `unq` tie breaking approach. The Δ% column reports the percent increase/decrease of the `RnkMsr` correlation with respect to the `Msr` one.

Measure	T08_30			T26_30			T27_30		
	Msr	RnkMsr	Δ%	Msr	RnkMsr	Δ%	Msr	RnkMsr	Δ%
P vs RBP_p05	0.7858	0.7858	+0.00%	0.8604	0.8604	+0.00%	0.7963	0.7963	+0.00%
P vs RR	0.7151	0.6322	-11.60%	0.8126	0.6049	-25.56%	0.7615	0.6764	-11.18%
P vs RBP_p03	0.7494	0.7858	+4.86%	0.8503	0.8604	+1.19%	0.7798	0.7963	+2.11%
P vs RBP_p08	0.8641	0.8447	-2.24%	0.9009	0.8944	-0.72%	0.8748	0.8465	-3.23%
P vs DCG_b02	0.9352	0.8962	-4.17%	0.9623	0.9312	-3.23%	0.9478	0.9203	-2.90%
P vs DCG_b10	0.9866	0.9243	-6.32%	0.9861	0.9254	-6.15%	0.9871	0.9360	-5.17%
RBP_p05 vs RBP_p03	0.9498	1.0000	+5.29%	0.9704	1.0000	+3.05%	0.9601	1.0000	+4.16%
RBP_p05 vs RBP_p08	0.9045	0.9082	+0.40%	0.9250	0.9351	+1.09%	0.9068	0.8998	-0.78%
RBP_p05 vs RR	0.8840	0.6733	-23.84%	0.9046	0.6228	-31.15%	0.9230	0.7590	-17.77%
RBP_p05 vs DCG_b02	0.8461	0.8166	-3.49%	0.8896	0.8593	-3.41%	0.8489	0.8082	-4.79%
RBP_p05 vs DCG_b10	0.7931	0.7686	-3.09%	0.8687	0.7958	-8.39%	0.8082	0.7886	-2.42%
RBP_p03 vs RBP_p08	0.8606	0.9082	+5.53%	0.9069	0.9351	+3.10%	0.8701	0.8998	+3.42%
RBP_p03 vs RR	0.9144	0.6733	-26.37%	0.9154	0.6228	-31.97%	0.9394	0.7590	-19.21%
RBP_p03 vs DCG_b02	0.8054	0.8166	+1.39%	0.8773	0.8593	-2.06%	0.8215	0.8082	-1.62%
RBP_p03 vs DCG_b10	0.7547	0.7686	+1.84%	0.8564	0.7958	-7.08%	0.7918	0.7886	-0.40%
RBP_p08 vs RR	0.8133	0.6430	-20.94%	0.8656	0.6307	-27.13%	0.8547	0.7213	-15.61%
RBP_p08 vs DCG_b02	0.9266	0.8822	-4.79%	0.9372	0.8939	-4.62%	0.9233	0.8646	-6.36%
RBP_p08 vs DCG_b10	0.8745	0.8333	-4.71%	0.9149	0.8304	-9.23%	0.8857	0.8434	-4.77%
RR vs DCG_b02	0.7668	0.5971	-22.13%	0.8375	0.5902	-29.52%	0.7998	0.6545	-18.17%
RR vs DCG_b10	0.7214	0.5896	-18.28%	0.8209	0.5475	-33.30%	0.7700	0.6514	-15.41%
DCG_b02 vs DCG_b10	0.9435	0.9273	-1.72%	0.9747	0.9221	-5.40%	0.9593	0.9632	+0.41%
R vs P	0.7948	1.0000	+25.82%	0.8901	1.0000	+12.35%	0.8191	1.0000	+22.08%
R vs RBP_p05	0.6848	0.7858	+14.76%	0.8102	0.8604	+6.20%	0.8098	0.7963	-1.67%
R vs RBP_p03	0.6648	0.7858	+18.20%	0.8052	0.8604	+6.86%	0.8169	0.7963	-2.52%
R vs RBP_p08	0.7424	0.8447	+13.78%	0.8391	0.8944	+6.59%	0.8082	0.8465	+4.74%
R vs RR	0.6611	0.6322	-4.37%	0.7841	0.6049	-22.86%	0.7971	0.6764	-15.14%
R vs DCG_b02	0.7756	0.8962	+15.54%	0.8730	0.9312	+6.67%	0.8169	0.9203	+12.67%
R vs DCG_b10	0.7957	0.9243	+16.15%	0.8853	0.9254	+4.54%	0.8184	0.9360	+14.37%
R vs AP	0.8859	0.8932	+0.83%	0.8709	0.9319	+7.01%	0.9078	0.9211	+1.46%
R vs nDCG_b02	0.7899	0.8962	+13.45%	0.8788	0.9312	+5.97%	0.8427	0.9203	+9.21%
R vs nDCG_b10	0.8338	0.9243	+10.85%	0.8911	0.9254	+3.86%	0.8490	0.9360	+10.25%
AP vs P	0.8244	0.8932	+8.35%	0.8583	0.9319	+8.58%	0.8481	0.9211	+8.61%
AP vs RBP_p05	0.7524	0.7936	+5.48%	0.8326	0.8413	+1.04%	0.8614	0.7980	-7.36%
AP vs RBP_p03	0.7271	0.7936	+9.14%	0.8160	0.8413	+3.09%	0.8575	0.7980	-6.94%
AP vs RBP_p08	0.8081	0.8587	+6.27%	0.8672	0.8730	+0.67%	0.8685	0.8466	-2.52%
AP vs RR	0.7030	0.5869	-16.52%	0.7761	0.5707	-26.47%	0.8249	0.6545	-20.66%
AP vs DCG_b02	0.8316	0.9673	+16.32%	0.8651	0.9632	+11.34%	0.8748	0.9757	+11.54%
AP vs DCG_b10	0.8280	0.9411	+13.67%	0.8672	0.9444	+8.90%	0.8575	0.9750	+13.69%
AP vs nDCG_b02	0.8454	0.9673	+14.42%	0.8709	0.9632	+10.60%	0.8912	0.9757	+9.49%
AP vs nDCG_b10	0.8563	0.9411	+9.90%	0.8687	0.9444	+8.72%	0.8849	0.9750	+10.17%
nDCG_b02 vs P	0.9330	0.8962	-3.95%	0.9579	0.9312	-2.79%	0.9305	0.9203	-1.10%
nDCG_b02 vs RBP_p05	0.8449	0.8166	-3.36%	0.8925	0.8593	-3.72%	0.8638	0.8082	-6.43%
nDCG_b02 vs RBP_p03	0.8051	0.8166	+1.42%	0.8773	0.8593	-2.06%	0.8395	0.8082	-3.73%
nDCG_b02 vs RBP_p08	0.9283	0.8822	-4.96%	0.9401	0.8939	-4.91%	0.9319	0.8646	-7.22%
nDCG_b02 vs RR	0.7684	0.5971	-22.30%	0.8389	0.5902	-29.65%	0.8179	0.6545	-19.98%
nDCG_b02 vs DCG_b02	0.9804	1.0000	+2.00%	0.9913	1.0000	+0.87%	0.9695	1.0000	+3.15%
nDCG_b02 vs DCG_b10	0.9418	0.9273	-1.54%	0.9719	0.9221	-5.12%	0.9429	0.9632	+2.16%
nDCG_b02 vs nDCG_b10	0.9368	0.9273	-1.01%	0.9690	0.9221	-4.84%	0.9577	0.9632	+0.57%
nDCG_b10 vs P	0.9568	0.9243	-3.40%	0.9832	0.9254	-5.88%	0.9525	0.9360	-1.73%
nDCG_b10 vs RBP_p05	0.7841	0.7686	-1.98%	0.8644	0.7958	-7.93%	0.8278	0.7886	-4.73%
nDCG_b10 vs RBP_p03	0.7472	0.7686	+2.86%	0.8535	0.7958	-6.76%	0.8129	0.7886	-2.99%
nDCG_b10 vs RBP_p08	0.8684	0.8333	-4.05%	0.9120	0.8304	-8.94%	0.8928	0.8434	-5.52%
nDCG_b10 vs RR	0.7212	0.5896	-18.25%	0.8194	0.5475	-33.18%	0.7881	0.6514	-17.35%
nDCG_b10 vs DCG_b02	0.9288	0.9273	-0.16%	0.9719	0.9221	-5.12%	0.9507	0.9632	+1.32%
nDCG_b10 vs DCG_b10	0.9610	1.0000	+4.06%	0.9899	1.0000	+1.02%	0.9616	1.0000	+3.99%

The correlation between RBP with $p = 0.3$ and RBP with $p = 0.5$ is 0.9498 while the correlation between their ranked versions is 1.0. As we discussed in Section VIII-A, RBP_p03 and RBP_p05 order runs in the same way and so their correlation should be 1.0. Therefore, this 5.3% underestimation of the similarity between them is just due to RBP_p03 not being an interval scale. Note that this case is somehow particularly severe since it induces us to attribute this 5.3% change to other reasons; typical explanations for such changes you may find in studies about

correlation among evaluation measures are: ''the user model behind RBP_p03 slightly differs from the RBP_p05 since it represents a more impatient or less motivated user'' or ''due to the smaller value of p, RBP_p03 is a slightly more top-heavy measure''; unfortunately, none of these explanations would be correct since this 5.3% change is just due to fact that the values of RBP_p03 are not equi-spaced, still ordering runs in exactly the same way as RBP_p05.

For the sake of completeness, we can observe that 0.9498 is the same correlation value reported in Table 2 between RBP_p03 and its ranked version. This is indeed correct since both RBP_p03 and RBP_p05 are mapped to the same ranked interval scale, which is just another mapping of the interval scale of RBP_p05; therefore, the correlation between RBP_p03 and its ranked version is the same as the correlation between RBP_p03 and RBP_p05.

Another interesting case is RR: its correlation with respect to P, RBP, and DCG is way over-estimated – in the range $12\%-26\%$ more on `T08_30`, $26\%-33\%$ more on `T26_30`, and $11\%-19\%$ more on `T27_30`, – mistakenly suggesting us that this measure is closer to the others much more than what it actually is. As it emerges from the previous discussion, RR is one of the measures which departs more from being an interval scale and which also has the highest number of tied values. Therefore, the computation of averages on RR and on its ranked version leads to sensibly different Rankings of Systems (RoS), as it is clearly shown in Table 2 when comparing RR to its ranked version. As a consequence, the correlation of the ranked version of RR with the other measures changes more than in the other cases.

2) MEASURES DEPENDING ON THE RECALL BASE

Before proceeding, a word of caution has to be made remembering that in this case our approach is just a surrogate, which improves the ''intervalness'' of a measure but stretches the steps of the scale. Therefore, all the increases/decreases in correlations should be taken as tendency to overestimation/underestimation rather than exact quantification of it.

Let us consider Precision and Recall: we know that on each topic they are the same interval scale and this is reflected in Table 4 in the correlation between their ranked version being 1.0. On the other hand, the correlation between the original measures tends to be underestimated by 26% on `T08_30`, 12% on `T26_30`, and 22% on `T27_30`. Apart from suggesting that these two measures should be considered closer to each other than what they usually are, this wide range of underestimation further stresses how much just the recall base can affect the averaging across topics and how careful we should be with such averages – not to say that we should avoid them at all.

Coherently with what we discussed in Section VIII-A2, we can observe that the correlation between DCG and nDCG tends to be underestimated by just $1\% - 4\%$, suggesting that they are practically equivalent. Therefore, even if usually nDCG is preferred over DCG because it is bounded and normalised, it could be actually better to use DCG instead, since it avoids issues with the recall base and it can be easily turned into a proper interval scale by using our transformation approach.

Let us now discuss AP with respect to Precision and Recall: the correlation with R is higher than the one with P and this is usually attributed to AP embedding the recall base in the same way as R does. However, when we turn to the ranked measures, we can see how the correlation between AP and R and AP and P is the same (for all the reasons already explained) and, especially, how this tends to be underestimated in the range $1\% - 9\%$, suggesting that AP is slightly closer to these two measures than what is usually thought.

Finally, let us consider AP with respect to DCG: their correlation tends to be underestimated in the range $9\% - 16\%$ and the correlation between their ranked versions is actually quite high, between 0.94 and 0.97. This suggests that, even if these two measures have very different formulations and the user model of DCG is considered much more realistic than the somehow artificial one of AP, when they are turned into their interval scale version, they are much closer than expected and that part of their difference could have been just due to their lack of ''intervalness''.

C. SIGNIFICANCE TESTING ANALYSIS. WHAT SYSTEMS ARE ACTUALLY DIFFERENT, OR NOT?

In this section, we analyse how the results of statistical significance tests change when using a measure or its ranked version. In other terms, we study how much statistical significance tests are impacted by using or not a proper interval scale. Indeed, as discussed in Section VIII-C, there are significance tests which assume to work with just an ordinal scale and others which assume to work with an interval scale and they should be somehow affected by using a measure which matches or not their assumptions. By impacted, we mean that we can observe some change in which systems are considered significantly different or not.

Moreover, as discussed in the previous sections, the recall base makes working across topics problematic at best and statistical significance tests typically perform some aggregation across topics. Therefore, they may be further affected by the recall base.

As described in Section VIII-C, we consider the following tests: Sign (ordinal scale assumption), Wilcoxon Rank Sum (ordinal scale assumption), Wilcoxon Signed Rank (interval scale assumption), Student's t (interval scale assumption), ANOVA (interval scale assumption), Kruskal-Wallis (ordinal scale assumption), and Friedman (ordinal scale assumption). For the ANOVA case we consider two alternatives:

- *One-Way ANOVA for the System Effect:* $y_{ij} = \mu_{..} + \alpha_j + \varepsilon_{ij}$ checks for the effect of α_j, $j = 1, \ldots, q$ different systems. It can be considered as an extension of the Student's t test to the comparison of multiple systems at the same time and it is the parametric counterpart of the Kruskal-Wallis Test.
- *Two-Way ANOVA for the Topic and System Effects:* A more accurate model $y_{ij} = \mu_{..} + \tau_i + \alpha_j + \varepsilon_{ij}$ which

accounts also for the effect of τ_i, $i = 1, \ldots, p$ topics, thus improving the estimation of the system effect as well. Note that this is the ANOVA model adopted by Tague- Sutcliffe and Blustein [106] and Banks et al. [12] when analysing TREC data. It is the parametric counterpart of the Friedman Test.

In the case of the ANOVA, Kruskal-Wallis, and Friedman tests we performed a Tukey Honestly Significant Difference (HSD) adjustment for multiple comparisons [58], [108].

Table 5 (measures not depending on the recall base) and Table 6 (measures depending on the recall base) show the results of the analyses in the case of the `T08_30`, `T26_30`, and `T27_30` tracks. Results for the other tracks are similar but not shown here for space reasons. For each test, the tables report:

- *Sig:* The total number of significantly different system pair using the original measure;
- *S2NS:* Number of pairs changed from significantly to not significantly different when passing from the original measure to its ranked version; within parentheses we report their ratio with respect to `Sig`;
- *NS2S:* Number of pairs changed from not significantly to significantly different when passing from the original measure to its ranked version; within parentheses we report their ratio with respect to `Sig`;
- $\Delta\%$: $\frac{\text{S2NS}+\text{NS2S}}{\text{Sig}}$ the ratio of the total number of pairs that changed significance when passing from the original measure to its ranked version.

In Tables 5 and 6 rows corresponding to significance tests based on an ordinal scale assumption are highlighted in grey.

In an ideal situation an oracle would have told us which pairs of systems are significantly different and which are not and this would have allowed us to exactly determine which pairs of systems were correctly detected by each measure and test. Unfortunately, this a priori knowledge is not available in practice. On the other hand, we are comparing a measure to its ranked version and we know that changes in the decision about what is significantly different and what is not are a consequence of the steps of the scale being rearranged from not equi-spaced and un-evenly distributed across the their range to equi-spaced and evenly distributed across their range. Therefore, we can interpret the `S2NS` count as a tendency to *false positives*, since it accounts for significantly different systems which are not significant when you remove the effect of uneven steps in the scale; in other terms, `S2NS` can be interpreted as a tendency of the ranked measure (the interval scale) to reduce *Type I errors*. Symmetrically, we can interpret the `NS2S` count as a tendency to *false negatives*, since it accounts for not significantly different systems which are significant when you remove the effect of uneven steps in the scale; in other terms, `NS2S` can be interpreted as a tendency of the ranked measure (the interval scale) to reduce *Type II errors*. Note that we are not claiming that an interval scale detects/removes false positives/negatives in any absolute sense; we are rather saying that, starting from whatever unknown level of false positive/negatives, we can interpret the `S2NS` and `NS2S` counts as a relative tendency to reduce false positives/negatives.

As a side note not regarding measures being interval scales or not, in Tables 5 and 6 we can observe that, as expected, parametric significance tests are more powerful than non-parametric ones, since they discover more significantly different pairs. Moreover, we can also observe as the Sign, Wilcoxon Rank Sum, Wilcoxon Signed Rank, and Student's t tests find many more significantly different system pairs than the ANOVA, Kruskal-Wallis, and Friedman tests. This increase is not due to more powerful tests but rather to the increase in Type I errors due to the lack of adjustment in multiple comparisons for the former tests. This further stresses the need for always adjusting for multiple comparisons, as also pointed out by Fuhr [49], Sakai [92].

We can observe from Tables 5 and 6, as very coarse and general trends, that the less close an evaluation measure is to be an interval scale, the stronger are the changes in statistical significance tests based on an interval scale assumption while those based on an ordinal scale assumption are not affected. On the other hand, the presence of the recall base generally affects both tests based on the ordinal scale assumption and those based on the interval scale assumption, being the latter more affected.

In particular, we have found that:

- for measures not depending on the recall base and significance test assuming an interval scale, we have an overall average increase in the `S2NS` count around 13% and in the `NS2S` around 5%. This suggests that the major impact is on reducing Type I errors still improving Type II errors and making the test more powerful;
- for measures depending on the recall base and significance test assuming an ordinal scale, we have an overall average increase in the `S2NS` count around 2% and in the `NS2S` around 4%. This suggests that there is a small reduction in Type I errors and some improvement in Type II errors, making the test a bit more powerful;
- for measures depending on the recall base and significance test assuming an interval scale, we have an overall average increase in the `S2NS` count around 10% and in the `NS2S` around 45%. This suggests that there is a sizeable reduction in Type I errors and a quite substantial improvement in Type II errors, making the test much more powerful.

In general, these results indicate that adopting a proper interval scale tends to reduce the Type I errors and, when the situation get more complicated because of the effect of the recall base across topics, it also brings substantially more power to the test.

Overall, if we consider the grand mean across all the tracks, measures, and significance test, we observe an overall change $\Delta\%$ around $25\% \pm 11\%$ in the decision about what is significantly different and what is not. Even without wishing to interpret it in term of Type I or Type II errors, this figure lets us understand the impact of using an interval scale or not, as well as the effect of the recall base.

TABLE 5. Measures not depending on the recall base: changes in significance test analyses between using a measure and its ranked version on tracks `T08_30`, `T26_30`, and `T27_30`, using the `unq` tie breaking approach.

P	T08_30 – 8256 system pairs				T26_30 – 2775 system pairs				T27_30 – 2556 system pairs			
	Sig	S2NS (%)	NS2S (%)	Δ%	Sig	S2NS (%)	NS2S (%)	Δ%	Sig	S2NS (%)	NS2S (%)	Δ%
Sign Test	5153	0 (0.00%)	0 (0.00%)	0.00%	1848	0 (0.00%)	0 (0.00%)	0.00%	1746	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4276	0 (0.00%)	0 (0.00%)	0.00%	1453	0 (0.00%)	0 (0.00%)	0.00%	1473	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5644	0 (0.00%)	0 (0.00%)	0.00%	2017	0 (0.00%)	0 (0.00%)	0.00%	1857	0 (0.00%)	0 (0.00%)	0.00%
Student's t Test	5633	0 (0.00%)	0 (0.00%)	0.00%	2007	0 (0.00%)	0 (0.00%)	0.00%	1830	0 (0.00%)	0 (0.00%)	0.00%
One-way ANOVA	1923	0 (0.00%)	0 (0.00%)	0.00%	454	0 (0.00%)	0 (0.00%)	0.00%	724	0 (0.00%)	0 (0.00%)	0.00%
Kruskal-Wallis Test	1740	0 (0.00%)	0 (0.00%)	0.00%	391	0 (0.00%)	0 (0.00%)	0.00%	692	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	3362	0 (0.00%)	0 (0.00%)	0.00%	1155	0 (0.00%)	0 (0.00%)	0.00%	1214	0 (0.00%)	0 (0.00%)	0.00%
Friedman Test	2595	0 (0.00%)	0 (0.00%)	0.00%	883	0 (0.00%)	0 (0.00%)	0.00%	925	0 (0.00%)	0 (0.00%)	0.00%
RBP_p05	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	4724	0 (0.00%)	0 (0.00%)	0.00%	1414	0 (0.00%)	0 (0.00%)	0.00%	1613	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4302	0 (0.00%)	0 (0.00%)	0.00%	1328	0 (0.00%)	0 (0.00%)	0.00%	1545	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5108	0 (0.00%)	0 (0.00%)	0.00%	1683	0 (0.00%)	0 (0.00%)	0.00%	1679	0 (0.00%)	0 (0.00%)	0.00%
Student's t Test	5091	0 (0.00%)	0 (0.00%)	0.00%	1711	0 (0.00%)	0 (0.00%)	0.00%	1635	0 (0.00%)	0 (0.00%)	0.00%
One-way ANOVA	1945	0 (0.00%)	0 (0.00%)	0.00%	377	0 (0.00%)	0 (0.00%)	0.00%	660	0 (0.00%)	0 (0.00%)	0.00%
Kruskal-Wallis Test	1678	0 (0.00%)	0 (0.00%)	0.00%	326	0 (0.00%)	0 (0.00%)	0.00%	680	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	2861	0 (0.00%)	0 (0.00%)	0.00%	733	0 (0.00%)	0 (0.00%)	0.00%	927	0 (0.00%)	0 (0.00%)	0.00%
Friedman Test	2228	0 (0.00%)	0 (0.00%)	0.00%	588	0 (0.00%)	0 (0.00%)	0.00%	806	0 (0.00%)	0 (0.00%)	0.00%
RBP_p03	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	4724	0 (0.00%)	0 (0.00%)	0.00%	1414	0 (0.00%)	0 (0.00%)	0.00%	1613	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4302	0 (0.00%)	0 (0.00%)	0.00%	1328	0 (0.00%)	0 (0.00%)	0.00%	1545	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5027	29 (0.58%)	110 (2.19%)	2.77%	1638	10 (0.61%)	55 (3.36%)	3.97%	1651	3 (0.18%)	31 (1.88%)	2.06%
Student's t Test	4801	70 (1.46%)	360 (7.50%)	8.96%	1551	19 (1.23%)	179 (11.54%)	12.77%	1486	7 (0.47%)	156 (10.50%)	10.97%
One-way ANOVA	1730	18 (1.04%)	233 (13.47%)	14.51%	317	3 (0.95%)	63 (19.87%)	20.82%	606	4 (0.66%)	58 (9.57%)	10.23%
Kruskal-Wallis Test	1678	0 (0.00%)	0 (0.00%)	0.00%	326	0 (0.00%)	0 (0.00%)	0.00%	680	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	2432	8 (0.33%)	437 (17.97%)	18.30%	573	2 (0.35%)	162 (28.27%)	28.62%	807	0 (0.00%)	120 (14.87%)	14.87%
Friedman Test	2228	0 (0.00%)	0 (0.00%)	0.00%	588	0 (0.00%)	0 (0.00%)	0.00%	806	0 (0.00%)	0 (0.00%)	0.00%
RBP_p08	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5105	0 (0.00%)	0 (0.00%)	0.00%	1615	0 (0.00%)	0 (0.00%)	0.00%	1678	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4375	0 (0.00%)	0 (0.00%)	0.00%	1378	0 (0.00%)	0 (0.00%)	0.00%	1487	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5508	165 (3.00%)	114 (2.07%)	5.07%	1879	70 (3.73%)	30 (1.60%)	5.32%	1846	110 (5.96%)	3 (0.16%)	6.12%
Student's t Test	5425	221 (4.07%)	123 (2.27%)	6.34%	1924	144 (7.48%)	60 (3.12%)	10.60%	1853	236 (12.74%)	14 (0.76%)	13.49%
One-way ANOVA	2183	200 (9.16%)	44 (2.02%)	11.18%	469	59 (12.58%)	15 (3.20%)	15.78%	712	98 (13.76%)	1 (0.14%)	13.90%
Kruskal-Wallis Test	1781	0 (0.00%)	0 (0.00%)	0.00%	384	0 (0.00%)	0 (0.00%)	0.00%	691	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	3474	257 (7.40%)	20 (0.58%)	7.97%	950	104 (10.95%)	21 (2.21%)	13.16%	1105	187 (16.92%)	6 (0.54%)	17.47%
Friedman Test	2589	0 (0.00%)	0 (0.00%)	0.00%	741	0 (0.00%)	0 (0.00%)	0.00%	851	0 (0.00%)	0 (0.00%)	0.00%
DCG_b02	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5227	0 (0.00%)	0 (0.00%)	0.00%	1781	0 (0.00%)	0 (0.00%)	0.00%	1738	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4425	0 (0.00%)	0 (0.00%)	0.00%	1457	0 (0.00%)	0 (0.00%)	0.00%	1499	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5743	260 (4.53%)	147 (2.56%)	7.09%	2020	127 (6.29%)	40 (1.98%)	8.27%	1887	91 (4.82%)	21 (1.11%)	5.94%
Student's t Test	5664	408 (7.20%)	161 (2.84%)	10.05%	2026	176 (8.69%)	47 (2.32%)	11.01%	1881	155 (8.24%)	15 (0.80%)	9.04%
One-way ANOVA	2125	633 (29.79%)	59 (2.78%)	32.56%	495	113 (22.83%)	3 (0.61%)	23.43%	734	153 (20.84%)	2 (0.27%)	21.12%
Kruskal-Wallis Test	1812	0 (0.00%)	0 (0.00%)	0.00%	400	0 (0.00%)	0 (0.00%)	0.00%	695	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	3609	601 (16.65%)	25 (0.69%)	17.35%	1129	179 (15.85%)	36 (3.19%)	19.04%	1201	328 (27.31%)	4 (0.33%)	27.64%
Friedman Test	2706	0 (0.00%)	0 (0.00%)	0.00%	876	0 (0.00%)	0 (0.00%)	0.00%	912	0 (0.00%)	0 (0.00%)	0.00%
DCG_b10	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5095	0 (0.00%)	0 (0.00%)	0.00%	1813	0 (0.00%)	0 (0.00%)	0.00%	1755	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4307	0 (0.00%)	0 (0.00%)	0.00%	1458	0 (0.00%)	0 (0.00%)	0.00%	1485	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	5663	260 (4.59%)	127 (2.24%)	6.83%	2019	105 (5.20%)	32 (1.58%)	6.79%	1861	60 (3.22%)	35 (1.88%)	5.10%
Student's t Test	5636	370 (6.56%)	132 (2.34%)	8.91%	2021	169 (8.36%)	50 (2.47%)	10.84%	1843	114 (6.19%)	36 (1.95%)	8.14%
One-way ANOVA	1958	747 (38.15%)	37 (1.89%)	40.04%	468	86 (18.38%)	12 (2.56%)	20.94%	727	187 (25.72%)	0 (0.00%)	25.72%
Kruskal-Wallis Test	1750	0 (0.00%)	0 (0.00%)	0.00%	390	0 (0.00%)	0 (0.00%)	0.00%	687	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	3438	623 (18.12%)	41 (1.19%)	19.31%	1139	130 (11.41%)	36 (3.16%)	14.57%	1216	256 (21.05%)	4 (0.33%)	21.38%
Friedman Test	2637	0 (0.00%)	0 (0.00%)	0.00%	882	0 (0.00%)	0 (0.00%)	0.00%	925	0 (0.00%)	0 (0.00%)	0.00%
RR	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	4074	0 (0.00%)	0 (0.00%)	0.00%	1305	0 (0.00%)	0 (0.00%)	0.00%	1309	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	3951	0 (0.00%)	0 (0.00%)	0.00%	1198	0 (0.00%)	0 (0.00%)	0.00%	1280	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Signed Rank Test	4578	479 (10.46%)	362 (7.91%)	18.37%	1497	201 (13.43%)	180 (12.02%)	25.45%	1462	94 (6.43%)	111 (7.59%)	14.02%
Student's t Test	4582	903 (19.71%)	454 (9.91%)	29.62%	1517	455 (29.99%)	180 (11.87%)	41.86%	1454	202 (13.89%)	110 (7.57%)	21.46%
One-way ANOVA	1691	288 (17.03%)	162 (9.58%)	26.61%	285	203 (71.23%)	2 (0.70%)	71.93%	696	92 (13.22%)	148 (21.26%)	34.48%
Kruskal-Wallis Test	1500	0 (0.00%)	0 (0.00%)	0.00%	218	0 (0.00%)	0 (0.00%)	0.00%	633	0 (0.00%)	0 (0.00%)	0.00%
Two-way ANOVA	2233	511 (22.88%)	131 (5.87%)	28.75%	527	349 (66.22%)	23 (4.36%)	70.59%	868	145 (16.71%)	92 (10.60%)	27.30%
Friedman Test	1813	0 (0.00%)	0 (0.00%)	0.00%	428	0 (0.00%)	0 (0.00%)	0.00%	739	0 (0.00%)	0 (0.00%)	0.00%

As in the case of the correlation analysis, these observations open some questions about IR experimentation: since violating the scale assumptions has an impact on the number of significant/not-significant detected pairs and on Type I and Type II errors, when we compare systems and algorithms, how much of the observed differences is just due to the scale violation bias? How many false positives/negatives are we observing? How much have the findings in the literature been affected by these phenomena?

1) MEASURES NOT DEPENDING ON THE RECALL BASE

Let us start from Precision and RBP_p05 in Table 5. As we already know, both of them are interval scales and, as expected, we do not observe any changes in using them or their ranked version.

As in the case of the correlation, we can take them as an example of *meaningful statements* in IR, since a statement like "There are 1,923 significantly different system pairs for Precision according to one-way ANOVA on `T08_30`" does not change its truth value for a permissible transformation of the scale.

As said, RBP_p03 orders systems in the same way as RBP_p05 but it is no more an interval scale. Coherently with this, we can see how the significance tests assuming just an ordinal case detect the same number of significantly different pairs for both RBP_p03 and RBP_p05. On the other hand,

TABLE 6. Measures depending on the recall base: changes in significance test analyses between using a measure and its ranked version on tracks `T08_30`, `T26_30`, and `T27_30`, using the `unq` tie breaking approach.

	T08_30 – 8256 system pairs				T26_30 – 2775 system pairs				T27_30 – 2556 system pairs			
R	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5153	0 (0.00%)	0 (0.00%)	0.00%	1848	0 (0.00%)	0 (0.00%)	0.00%	1746	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	3494	14 (0.40%)	796 (22.78%)	23.18%	1239	25 (2.02%)	239 (19.29%)	21.31%	1254	17 (1.36%)	236 (18.82%)	20.18%
Wilcoxon Signed Rank Test	5434	185 (3.40%)	395 (7.27%)	10.67%	1892	35 (1.85%)	160 (8.46%)	10.31%	1799	27 (1.50%)	85 (4.72%)	6.23%
Student's t Test	5073	303 (5.97%)	863 (17.01%)	22.98%	1723	62 (3.60%)	346 (20.08%)	23.68%	1554	34 (2.19%)	310 (19.95%)	22.14%
One-way ANOVA	409	0 (0.00%)	1514 (370.17%)	370.17%	69	0 (0.00%)	385 (557.97%)	557.97%	574	3 (0.52%)	153 (26.66%)	27.18%
Kruskal-Wallis Test	1417	28 (1.98%)	351 (24.77%)	26.75%	259	0 (0.00%)	132 (50.97%)	50.97%	653	16 (2.45%)	55 (8.42%)	10.87%
Two-way ANOVA	2440	47 (1.93%)	969 (39.71%)	41.64%	683	3 (0.44%)	475 (69.55%)	69.99%	923	15 (1.63%)	306 (33.15%)	34.78%
Friedman Test	2595	0 (0.00%)	0 (0.00%)	0.00%	883	0 (0.00%)	0 (0.00%)	0.00%	925	0 (0.00%)	0 (0.00%)	0.00%
AP	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5233	0 (0.00%)	0 (0.00%)	0.00%	1801	0 (0.00%)	0 (0.00%)	0.00%	1727	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4067	25 (0.61%)	385 (9.47%)	10.08%	1455	84 (5.77%)	85 (5.84%)	11.62%	1446	48 (3.32%)	96 (6.64%)	9.96%
Wilcoxon Signed Rank Test	5659	242 (4.28%)	169 (2.99%)	7.26%	1896	91 (4.80%)	128 (6.75%)	11.55%	1853	97 (5.23%)	38 (2.05%)	7.29%
Student's t Test	5156	339 (6.57%)	559 (10.84%)	17.42%	1683	145 (8.62%)	336 (19.96%)	28.58%	1759	118 (6.71%)	109 (6.20%)	12.91%
One-way ANOVA	391	3 (0.77%)	955 (244.25%)	245.01%	59	0 (0.00%)	295 (500.00%)	500.00%	529	60 (11.34%)	65 (12.29%)	23.63%
Kruskal-Wallis Test	1593	19 (1.19%)	225 (14.12%)	15.32%	380	20 (5.26%)	37 (9.74%)	15.00%	700	21 (3.00%)	11 (1.57%)	4.57%
Two-way ANOVA	2508	171 (6.82%)	560 (22.33%)	29.15%	636	31 (4.87%)	360 (56.60%)	61.48%	913	105 (11.50%)	73 (8.00%)	19.50%
Friedman Test	2706	0 (0.00%)	0 (0.00%)	0.00%	883	0 (0.00%)	0 (0.00%)	0.00%	915	0 (0.00%)	0 (0.00%)	0.00%
nDCG_b02	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5227	0 (0.00%)	0 (0.00%)	0.00%	1781	0 (0.00%)	0 (0.00%)	0.00%	1738	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4420	49 (1.11%)	54 (1.22%)	2.33%	1457	14 (0.96%)	14 (0.96%)	1.92%	1530	52 (3.40%)	21 (1.37%)	4.77%
Wilcoxon Signed Rank Test	5726	245 (4.28%)	149 (2.60%)	6.88%	2004	115 (5.74%)	44 (2.20%)	7.93%	1892	96 (5.07%)	21 (1.11%)	6.18%
Student's t Test	5623	384 (6.83%)	178 (3.17%)	9.99%	2013	174 (8.64%)	58 (2.88%)	11.53%	1883	163 (8.66%)	21 (1.12%)	9.77%
One-way ANOVA	2159	667 (30.89%)	59 (2.73%)	33.63%	498	116 (23.29%)	3 (0.60%)	23.90%	820	237 (28.90%)	0 (0.00%)	28.90%
Kruskal-Wallis Test	1827	34 (1.86%)	19 (1.04%)	2.90%	418	20 (4.78%)	2 (0.48%)	5.26%	718	25 (3.48%)	2 (0.28%)	3.76%
Two-way ANOVA	3642	631 (17.33%)	22 (0.60%)	17.93%	1120	172 (15.36%)	38 (3.39%)	18.75%	1201	330 (27.48%)	6 (0.50%)	27.98%
Friedman Test	2706	0 (0.00%)	0 (0.00%)	0.00%	876	0 (0.00%)	0 (0.00%)	0.00%	912	0 (0.00%)	0 (0.00%)	0.00%
nDCG_b10	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**	**Sig**	**S2NS (%)**	**NS2S (%)**	**Δ%**
Sign Test	5095	0 (0.00%)	0 (0.00%)	0.00%	1813	0 (0.00%)	0 (0.00%)	0.00%	1755	0 (0.00%)	0 (0.00%)	0.00%
Wilcoxon Rank Sum Test	4312	52 (1.21%)	47 (1.09%)	2.30%	1453	15 (1.03%)	20 (1.38%)	2.41%	1532	62 (4.05%)	15 (0.98%)	5.03%
Wilcoxon Signed Rank Test	5632	250 (4.44%)	148 (2.63%)	7.07%	2021	111 (5.49%)	36 (1.78%)	7.27%	1851	65 (3.51%)	50 (2.70%)	6.21%
Student's t Test	5579	370 (6.63%)	189 (3.39%)	10.02%	2012	162 (8.05%)	52 (2.58%)	10.64%	1816	122 (6.72%)	71 (3.91%)	10.63%
One-way ANOVA	2012	814 (40.46%)	50 (2.49%)	42.94%	480	97 (20.21%)	11 (2.29%)	22.50%	830	290 (34.94%)	0 (0.00%)	34.94%
Kruskal-Wallis Test	1782	50 (2.81%)	18 (1.01%)	3.82%	397	7 (1.76%)	0 (0.00%)	1.76%	728	44 (6.04%)	3 (0.41%)	6.46%
Two-way ANOVA	3478	670 (19.26%)	48 (1.38%)	20.64%	1125	121 (10.76%)	41 (3.64%)	14.40%	1203	281 (23.36%)	42 (3.49%)	26.85%
Friedman Test	2637	0 (0.00%)	0 (0.00%)	0.00%	882	0 (0.00%)	0 (0.00%)	0.00%	925	0 (0.00%)	0 (0.00%)	0.00%

significance tests assuming an interval scale are affected by this difference between RBP_p03 and RBP_p05, causing an overall change $\Delta\%$ in the range $2\% - 29\%$. In particular, we observe a marked increase in the number of significantly different pairs (`NS2S` up to 28%), i.e. the reduction in the number of false negatives, and a very marginal decrease in the number of not significantly different ones (`S2NS` around 1%), i.e. the reduction in the number of false positives. In the case of RBP_p08 we note a much more marked increase in the number of not significantly different pairs (`S2NS` up to 17%), i.e. a reduction in the number of false positives; on the other hand, the increase in the number of significantly different pairs (`NS2S` around 2%), i.e. the reduction in the number of false negatives, is more marginal.

Why do we observe such a different behaviour between RBP_p03 and RBP_p08? If we look at Figure 7, we can see that RBP_p03 condenses values at the top and the bottom of the range of possible values, in spans with a very small range of values but containing the same number of runs. As a consequence, when the ranked version of RBP_p03 equispaces these values, runs that before were very close, and possibly not significantly different (`NS`) in the ranked version become more distant, and possibly significantly different (`S`); and this can explain why the `NS2S` case is more prominent for RBP_p03. On the other hand, RBP_p08 uses all the possible range of values but very few runs, roughly 20% packed at the bottom and at the top, cover almost 50% of the range of values while the remaining 80% of the runs, in the middle part, cover the other 50% of the range. Therefore, when we pass to the ranked version of the measure, very few runs which were very distant, and possibly significantly different (`S`), become closer, and possibly not significantly different (`NS`); viceversa, many runs which were very close, and possibly not significantly different (`NS`), may become a little bit more distant, and possibly (but not necessarily) significantly different (`S`). As a consequence, the effect on `S2NS` is more prominent than the one on `NS2S`.

In the case of DCG we observe a behaviour similar to the one of RBP_p08, being the increase on `S2NS` and the reduction in the false positives even more prominent. If we look at Figure 8, we can see how DCG is sharper than RBP_p08 both at the top and the bottom of the range – less than 10% of the runs account for almost 50% of the range of values – making even fewer runs falling more apart.

Finally, RR exhibits both effects: a very remarkable increase in `S2NS`, i.e. a reduction in the false positives, and a sizeable increase in `NS2S`, i.e. a reduction in the false negatives, causing an overall change $\Delta\%$ up to a 72%. If we consider Figure 6, we can see how most of the runs, over 90%, are concentrated in just 4 possible values which are quite distant, possibly making them significantly different (`S`); when we move to the ranked version, these 4 values become much closer, possibly making the runs not significantly different (`NS`); and this explains the big `S2NS` effect. Vice versa, few runs, less than 10%, account for just the 20% of the range of values in the lower quartile; when we move to the ranked version, these values become more distant, possibly making the runs significantly different (`S`); since this change may

affect a smaller number of runs, this explains why NS2S tends to be more moderate with respect to S2NS.

2) MEASURES DEPENDING ON THE RECALL BASE

As in the case of the correlation among measures, a word of caution has to be made remembering that in the case of measures depending on the recall base our approach is just a surrogate, which improves the ''intervalness'' of a measure but stretches the steps of the scale. Therefore, all the changes in the significantly different system pairs should be taken as tendency rather than exact quantification.

Regarding the significance tests assuming an ordinal scale in Table 6, we can note – with the exception of the Sign and Friedman tests – that these tests are also affected by the transformation to an interval scale for an overall change $\Delta\%$ of up to 51%. This further confirms that aggregating across topics when the recall base changes can cause variations which go well beyond the loss of ''intervalness''.

As another general trend, we can see that significance tests based on an interval scale assumption are generally more affected, since they experience both the violation of their assumptions and the effect of the recall base.

If we consider Recall, we can see how the most prominent effect is the underestimation of significant differences with a very big increase in the number of significantly different pairs (NS2S), i.e. a reduction in the number of false positives, up to a striking 558% for the one-way ANOVA. Considering that the interval scale behind Recall is the same as the one behind Precision, these figures tell us how big the loss of power is for Recall, mostly due to the impact of aggregating across topics with different recall bases.

In the case of nDCG we can observe a behaviour quite similar to one of DCG, with an overall change $\Delta\%$ just a bit bigger than the one of DCG. Considering that both DCG and nDCG share the same interval scale, this further suggests to use DCG to avoid the further bias due to the recall base.

IX. CONCLUSION AND FUTURE WORK

We have discussed the problem that IR evaluation measures typically are not interval scales. In a strict sense, according to Stevens's prescriptions, this fact has several consequences: you should neither compute means, variances, and confidence intervals nor perform statistical significance tests which assume an interval scale. We have provided a detailed discussion on the motivations and needs behind the interval scales, both in the general field of the representational theory of measurement and in the IR context in particular. Moreover, since these prescriptions are very much debated both within and outside IR we have presented viewpoints and opinions both supporting and opposing these ''prohibitions''. It is a matter of fact that these ''prohibitions'' have not been accurately investigated by the IR community yet. However, the validity of experimental results and drawn conclusions should be considered ''at risk'' when departing and violating assumptions, at least until when the impact of these violations has been thoroughly investigated, as it has been happening so far in IR.

The main motivation for IR measures not being interval scales is that their values are not equi-spaced. Therefore, we have proposed a straightforward yet powerful way to turn any measure into an interval scale by considering how all the possible runs are ranked by the measure and keeping the unique ranks, i.e. after removing tied values, as values of the mapped measures. These ranks are equi-spaced by construction and preserve the same order of runs of the original measure. In this way, we obtain an interval scale able to represent the order of runs produced by the original measure.

We have also shown that the situation in IR is worsened by the fact that mixing runs of different length and different recall bases for different topics actually means mixing different scales, being them interval or not. Therefore, computing aggregations across runs and topics in such a way can lead to invalid results. While the run length issue can be mitigated by ensuring that all the runs have the same length, the recall base one is more problematic, since you cannot force a single recall base for all the topics. Therefore, this discourages the use of measures depending on the recall base.

Overall, this discussion led us to raise the fundamental question that IR should be more concerned with being able to rely on *meaningful statements*, i.e. statements whose truth values do not change when you perform legitimate transformations of the underlying scale, since they ensure for more valid and generalisable inferences.

Relying on several TREC collections, we have conducted a thorough experimentation on several (popular) state-of-the-art evaluation measures in order to assess the differences between using an evaluation measure and its interval-scaled version.

The correlation analysis has shown that the relationship between evaluation measures and their interval-scaled version matches the expected theoretical properties and that not using an interval scale somehow inflates the differences among evaluation measures. Notably, RR represents an exception since its departure from being an interval scale makes it look to be more similar to other measures than what it actually is.

Most importantly, the correlation analysis provides us with a rough estimator of how much interval scale an evaluation measure is and it represents the first attempt to quantify how much evaluation measures depart from their assumptions.

The analysis on many different types of statistical significance tests has clearly shown the impact of passing from an evaluation measure to its interval-scaled version. In particular, for measures not depending on the recall base, the transformation provides benefits in terms of reduced Type I error and some increase in power of the test. While for measures depending on the recall base, it produces sizeable improvements in terms of Type II error and power of the test, still delivering substantial enhancements in terms of Type I error. Even apart from any interpretation in terms

of Type I and Type II errors, we observed an overall mean change around 25% in the decision about which systems are significantly different and which are not.

Our results on both the correlation analysis and the statistical significance tests open the question about which claims and findings in the IR literature would be impacted by these difference or, in other terms, which statements made in IR so far would be actually *meaningful*. This should not be intended as decades of IR research have drawn wrong conclusions but rather that we should strengthen the foundations of our field, by further validating our (previous) results and by more explicitly stating the limitations and under which conditions conclusions are drawn.

The main limitation of the proposed approach is practical, since first you need to generate all the possible runs of REL^N and then you have compute the evaluation measures on all these runs. For increasing values of N, and even more in the case of multi-graded relevance, this becomes practically infeasible. Therefore, our future work will concern approximating this generation process in order to make it possible to deal with runs of whatever length.

REFERENCES

[1] R. P. Abelson and J. W. Tukey, "Efficient conversion of non-metric information into metric information," in *Proc. Social Statist. Sect. Amer. Stat. Assoc.* Washington, DC, USA: American Statistical Association, 1959, pp. 226–230.

[2] E. W. Adams, R. F. Fagot, and R. E. Robinson, "A theory of appropriate statistics," *Psychometrika*, vol. 30, no. 2, pp. 99–127, Jun. 1965.

[3] J. Allan, D. K. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees, "TREC 2017 common core track overview," in *Proc. 26th Text Retr. Conf. (TREC)*. E. M. Voorhees and A. Ellis, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 2018, pp. 1–14.

[4] J. Allan, D. K. Harman, E. Kanoulas, and E. M. Voorhees, "TREC 2018 common core track overview," in *Proc. 27th Text Retr. Conf. (TREC)*, E. M. Voorhees and A. Ellis, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 2019, pp. 1–10.

[5] E. Amigó and S. Mizzaro, "On the nature of information access evaluation metrics: A unifying framework," *Inf. Retr. J.*, vol. 23, no. 3, pp. 318–386, Jun. 2020.

[6] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Inf. Retr.*, vol. 12, no. 4, pp. 461–486, 2009.

[7] E. Amigó, J. Gonzalo, and F. Verdejo, "A general evaluation measure for document organization tasks," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 643–652.

[8] E. Amigó, D. Spina, and J. Carrillo-de-Albornoz, "An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 625–634.

[9] E. Amigó, J. Gonzalo, F. Verdejo, and D. Spina, "A comparison of filtering evaluation metrics based on formal constraints," *Inf. Retr. J.*, vol. 22, no. 6, pp. 581–619, Dec. 2019.

[10] N. H. Anderson, "Scales and statistics: Parametric and nonparametric," *Psychol. Bull.*, vol. 58, no. 4, pp. 305–316, Jul. 1961.

[11] K. Balog, "Expertise retrieval," *Found. Trends Inf. Retr.*, vol. 6, nos. 2–3, pp. 127–256, 2012.

[12] D. Banks, P. Over, and N.-F. Zhang, "Blind men and elephants: Six approaches to TREC data," *Inf. Retr.*, vol. 1, nos. 1–2, pp. 7–34, May 1999.

[13] P. Bollmann, "Two axioms for evaluation measures in information retrieval," in *Proc. 3rd Joint BCS ACM Symp. Res. Develop. Inf. Retr.*, C. J. van Rijsbergen, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1984, pp. 233–245.

[14] P. Bollmann and V. S. Cherniavsky, "Measurement-theoretical investigation of the MZ-metric," in *Proc. 3rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, C. J. van Rijsbergen, Ed. New York, NY, USA: ACM Press, 1980, pp. 256–267.

[15] P. Bollmann and V. S. Cherniavsky, "Restricted evaluation in information retrieval," in *Proc. 4th Annu. Int. ACM SIGIR Conf. Inf. Storage Retr. Theor. Issues Inf. Retr. (SIGIR)*, 1981, pp. 15–21.

[16] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proc. 27th Annu. Int. Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2004, pp. 25–32.

[17] C. Hauff and F. de Jong, "Retrieval system evaluation," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2010, pp. 53–78.

[18] S. Burnett, S. Clarke, M. Davis, R. Edwards, and A. Kellett, *Enterprise Search and Retrieval. Unlocking the Organisation's Potential*. London, U.K.: Butler Direct Limited, Oct. 2006.

[19] L. Busin and S. Mizzaro, "Axiometrics: An axiomatic approach to information retrieval effectiveness metrics," in *Proc. 4th Int. Conf. Theory Inf. Retr. (ICTIR)*, O. Kurland, D. Metzler, C. Lioma, B. Larsen, and P. Ingwersen, Eds. New York, NY, USA: ACM Press, 2013, pp. 22–29.

[20] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff, "Reliable information retrieval evaluation with incomplete and biased judgements," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2007, pp. 63–70.

[21] S. Büttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge, MA, USA: MIT Press, 2010.

[22] N. R. Campbell, *Physics: The Elements*. Cambridge, U.K.: Cambridge Univ. Press, 1920.

[23] N. R. Campbell, *An Account of the Principles of Measurement and Calculation*. London, U.K.: Longmans, Green, 1928.

[24] D. Carmel and E. Yom-Tov, *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool: San Rafael, CA, USA, 2010.

[25] B. Carterette, "System effectiveness, user models, and user utility: A conceptual framework for investigation," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. (SIGIR)*, 2011, pp. 903–912.

[26] B. A. Carterette, "Multiple testing in statistical analysis of systems-based information retrieval experiments," *ACM Trans. Inf. Syst.*, vol. 30, no. 1, pp. 1–34, Feb. 2012.

[27] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 651–658.

[28] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 621–630.

[29] T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani, Eds., in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*. New York, NY, USA: ACM Press, 2008.

[30] C. Cleverdon, "The cranfield tests on index languages devices," *Aslib Proc.*, vol. 19, no. 6, pp. 173–194, Jun. 1967.

[31] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1998, pp. 282–289.

[32] P. C. Cozby and S. C. Bates, *Methods in Behavioral Research*, 13th ed. New York, NY, USA: McGraw-Hill, 2018.

[33] W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds., in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*. New York, NY, USA: ACM Press, 1998.

[34] B. A. Davey and H. A. Priestley, *Introduction to Lattices Order*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[35] *European Commission. Commission Regulation (EC) No 607/2009 of 14 July 2009 Laying Down Certain Detailed Rules for the Implementation of Council Regulation (EC) No 479/2008 as Regards Protected Designations of Origin and Geographical Indications, Traditional Terms, Labelling and Presentation of Certain Wine Sector Products*, document OJ L 193, 24.7.2009, 52:60–139, Official Journal of the European Union, Jul. 2009.

[36] *European Commission. Commission Delegated Regulation (EC) No 2019/33 of 17 October 2018 Supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as Regards Applications for Protection of Designations of Origin, Geographical Indications and Traditional Terms in the Wine Sector, the Objection Procedure, Restrictions of Use, Amendments to Product Specifications, Cancellation of Protection, and Labelling and Presentation*, document OJ L 9, 11.1.2019, 62:2–45, Official Journal of the European Union, Jan. 2019.

[37] J. C. Falmagne and L. Narens, "Scales and meaningfulness of quantitative laws," *Synthese*, vol. 55, no. 3, pp. 287–325, Jun. 1983.

[38] A. Ferguson, C. S. Myers, R. J. Bartlett, H. Banister, F. C. Bartlett, W. Brown, N. R. Campbell, K. J. W. Craik, J. Drever, J. Guild, R. A. Houstoun, J. O. Irwin, G. W. C. Kaye, S. J. F. Philpott, L. F. Richardson, J. H. Shaxby, T. Smith, R. H. Thouless, and W. S. Tucker, "Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events," *Advancement Sci.*, vol. 1, no. 2, pp. 331–349, 1940.

[39] M. Ferrante, N. Ferro, and M. Maistro, "Towards a formal framework for utility-oriented measurements of retrieval effectiveness," in *Proc. Int. Conf. Theory Inf. Retr.*, Sep. 2015, pp. 21–30.

[40] M. Ferrante, N. Ferro, and S. Pontarollo, "Are IR evaluation measures on an interval scale," in *Proc. 3rd ACM SIGIR Int. Conf. Theory Inf. Retr. (ICTIR)*, J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, Eds. New York, NY, USA: ACM Press, 2017, pp. 67–74.

[41] M. Ferrante, N. Ferro, and S. Pontarollo, "A general theory of IR evaluation measures," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 409–422, Mar. 2019.

[42] M. Ferrante, N. Ferro, and E. Losiouk, "How do interval scales help us with better understanding IR evaluation measures?" *Inf. Retr. J.*, vol. 23, no. 3, pp. 289–317, Jun. 2020.

[43] N. Ferro, "What does affect the correlation among evaluation measures?" *ACM Trans. Inf. Syst.*, vol. 36, no. 2, pp. 1–40, Sep. 2017.

[44] N. Ferro and C. Peters, Eds. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, vol. 41. Berlin, Germany: Springer The Information Retrieval Series, 2019.

[45] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh, U.K.: Oliver & Boyd, 1925.

[46] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.

[47] M. Friedman, "A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 34, no. 205, p. 109, Mar. 1939.

[48] N. Fuhr, "Salton award lecture: Information retrieval as engineering science," *SIGIR Forum*, vol. 46, no. 2, pp. 19–28, Dec. 2012.

[49] N. Fuhr, "Some common mistakes in IR evaluation, and how they can be avoided," *ACM SIGIR Forum*, vol. 51, no. 3, pp. 32–41, Feb. 2018.

[50] J. Gaito, "Non-parametric methods in psychological research," *Psychol. Rep.*, vol. 5, no. 1, pp. 115–125, Mar. 1959.

[51] J. Gaito, "Measurement scales and statistics: Resurgence of an old misconception," *Psychol. Bull.*, vol. 87, no. 3, pp. 564–567, 1980.

[52] P. L. Gardner, "Scales and statistics," *Rev. Educ. Res.*, vol. 45, no. 1, pp. 43–57, Winter 1975.

[53] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference*, 5th ed. London, U.K.: Chapman & Hall, 2011.

[54] D. J. Hand, "Statistics and the theory of measurement," *J. Roy. Stat. Society. Ser. Statist. Soc.*, vol. 159, no. 3, pp. 445–492, 1996.

[55] D. K. Harman, *Information Retrieval Evaluation*. Morgan & Claypool: San Rafael, CA, USA, 2011.

[56] D. K. Harman and E. M. Voorhees, Eds., *TREC. Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA: MIT Press, 2005.

[57] M. H. Heine, "Distance between sets as an objective measure of retrieval effectiveness," *Inf. Storage Retr.*, vol. 9, no. 3, pp. 181–198, Mar. 1973.

[58] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*. Hoboken, NJ, USA: Wiley, 1987.

[59] D. Hull, "Using statistical testing in the evaluation of retrieval experiments," in *Proc. 16th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1993, pp. 329–338.

[60] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.

[61] J. Jiang and J. Allan, "Correlation between system and user metrics in a session," in *Proc. ACM Conf. Hum. Inf. Interact. Retr.*, Mar. 2016, pp. 285–288.

[62] J. Kekäläinen and K. Järvelin, "Using graded relevance assessments in IR evaluation," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 53, no. 13, pp. 1120–1129, Nov. 2002.

[63] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," *Found. Trends Inf. Retr.*, vol. 3, no. 1—2, pp. 1–224, 2009.

[64] M. G. Kendall, *Rank Correlation Methods*. Oxford, U.K.: Griffin, 1948.

[65] D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement. Additive and Polynomial Representations*, vol. 1. New York, NY, USA: Academic, 1971.

[66] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

[67] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. New York, NY, USA: McGraw-Hill, 2005.

[68] F. M. Lord, "On the statistical treatment of football numbers," *Amer. Psychol.*, vol. 8, no. 12, pp. 750–751, Dec. 1953.

[69] D. E. Losada, J. Parapar, and Á. Barreiro, "Feeling lucky?: Multi-armed bandits for ordering judgements in pooling-based evaluation," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, Apr. 2016, pp. 1027–1034.

[70] R. D. Luce and J. W. Tukey, "Simultaneous conjoint measurement: A new type of fundamental measurement," *J. Math. Psychol.*, vol. 1, no. 1, pp. 1–27, Jan. 1964.

[71] R. D. Luce, D. H. Krantz, P. Suppes, and A. Tversky, *Foundations of Measurement. Representation, Axiomatization, and Invariance*, vol. 3. New York, NY, USA: Academic, 1990.

[72] M. Lupu and A. Hanbury, "Patent retrieval," *Found. Trends Inf. Retr.*, vol. 7, no. 1, pp. 1–97, 2013.

[73] E. Maddalena and S. Mizzaro, "Axiometrics: Axioms of information retrieval effectiveness metrics," in *Proc. 6th Int. Workshop Evaluating Inf. Access (EVIA)*, S. Mizzaro and R. Song, Eds. Tokyo, Japan: National Institute of Informatics, 2014, pp. 17–24.

[74] H. M. Marcus-Roberts and F. S. Roberts, "Meaningless statistics," *J. Educ. Statist.*, vol. 12, no. 4, pp. 383–394, Dec. 1987.

[75] J. Michell, "Measurement scales and statistics: A clash of paradigms," *Psychol. Bull.*, vol. 100, no. 3, pp. 398–407, 1986.

[76] J. Michel, *An Introduction to the Logic of Psychological Measurement*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1990.

[77] A. Moffat, "Seven numeric properties of effectiveness metrics," in *Proc. 9th Asia Inf. Retr. Societies Conf. (AIRS)* in Lecture Notes in Computer Science, vol. 8281, R. E. Banchs, F. Silvestri, T.-Y. Liu, M. Zhang, S. Gao, and J. Lang, Eds. Berlin, Germany: Springer, 2013, pp. 1–12.

[78] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Trans. Inf. Syst.*, vol. 27, no. 1, pp. 1–27, Dec. 2008.

[79] A. Moffat, P. Thomas, and F. Scholer, "Users versus models: What observation tells us about effectiveness metrics," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 659–668.

[80] H. Müller, "Medical (visual) information retrieval," in *Information Retrieval Meets Information Visualization—PROMISE Winter School 2012, Revised Tutorial Lectures* (Lecture Notes in Computer Science), vol. 7757, M. Agosti, N. Ferro, P. Forner, H. Müller, and G. Santucci, Eds. Berlin, Germany: Springer, 2013, pp. 155–166.

[81] L. Narens, *Theories Meaningfullness*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2002.

[82] R. O'Brien, "The relationship between ordinal measures and their underlying values: Why all the disagreement?" *Qual. Quantity*, vol. 19, no. 3, pp. 265–277, Jun. 1985.

[83] F. S. Roberts, "Applications of the theory of meaningfulness to psychology," *J. Math. Psychol.*, vol. 29, no. 3, pp. 311–332, Sep. 1985.

[84] S. Robertson, "On GMAP: And other transformations," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2006, pp. 78–83.

[85] B. R. Rowe, D. W. Wood, A. L. Link, and D. A. Simoni, *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*, document RTI Project Number 0211875, RTI Int., Raleigh, NC, USA, Jul. 2010. [Online]. Available: http://trec.nist.gov/pubs/2010.economic.impact.pdf

[86] T. Sakai, "Evaluating evaluation metrics based on the bootstrap," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2006, pp. 525–532.

[87] T. Sakai, "Evaluation with informational and navigational intents," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 499–508.

[88] T. Sakai, "Metrics, statistics, tests," in *Bridging Between Information Retrieval and Databases—PROMISE Winter School 2013* (Lecture Notes in Computer Science), vol. 8173, N. Ferropp, Ed. Berlin, Germany: Springer, 2014, pp. 116–163.

[89] T. Sakai, "Statistical reform in information retrieval," *SIGIR Forum*, vol. 48, no. 1, pp. 3–12, Jun. 2014.

[90] T. Sakai, "A simple and effective approach to score standardisation," in *Proc. ACM Int. Conf. Theory Inf. Retr.*, Sep. 2016, pp. 95–104.

[91] T. Sakai, "Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 5–14.
[92] T. Sakai, "On Fuhr's guideline for IR evaluation," *ACM SIGIR Forum*, vol. 54, no. 1, pp. 1–8, Jun. 2020.
[93] T. Sakai, D. W. Oard, and N. Kando, Eds., *Evaluating Information Retrieval and Access Tasks—NTCIR's Legacy of Research Impact* (The Information Retrieval Series), vol. 43. Berlin, Germany: Springer, 2021.
[94] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *J. ACM*, vol. 15, no. 1, pp. 8–36, Jan. 1968.
[95] M. Sanderson, "Test collection based evaluation of information retrieval systems," *Found. Trends Inf. Retr.*, vol. 4, no. 4, pp. 247–375, 2010.
[96] J. Sauro and J. R. Lewis, *Quantifying the User Experience: Practical Statistics for User Research*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2016.
[97] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Inf. Process. Manage.*, vol. 33, no. 4, pp. 495–512, Jul. 1997.
[98] A. Z. Scholten and D. Borsboom, "A reanalysis of Lord's statistical treatment of football numbers," *J. Math. Psychol.*, vol. 53, no. 2, pp. 69–75, Apr. 2009.
[99] F. Sebastiani, "Evaluation measures for quantification: An axiomatic approach," *Inf. Retr. J.*, vol. 23, no. 3, pp. 255–288, Jun. 2020.
[100] V. L. Senders, *Measurement and Statistics: A Basic Text Emphasizing Behavioral Science Applications*. New York, NY, USA: Oxford Univ. Press, 1958.
[101] S. Siegel, *Nonparametric Statistics: For the Behavioral Science*. New York, NY, USA: McGraw-Hill, 1956.
[102] A. Singhal, J. Choi, D. Hindle, and F. C. N. Pereira, "AT&T at TREC-6: SDR track," in *Proc. 6th Text Retr. Conf. (TREC)*, E. M. Voorhees and D. K. Harman, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 1997, pp. 227–232.
[103] S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.
[104] W. S. Gosset, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.
[105] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky, *Foundations of Measurement. Geometrical, Threshold, and Probabilistic Representations*, vol. 2. New York, NY, USA: Academic, 1989.
[106] J. M. Tague-Sutcliffe and J. Blustein, "A statistical Analysis of the TREC-3 Data," in *Proc. 3rd Text Retr. Conf. (TREC)*, D. K. Harman, Ed. Washington, DC, USA: National Institute of Standards and Technology (NIST), 1994, pp. 385–398.
[107] J. T. Townsend and F. G. Ashby, "Measurement scales and statistics: The misconception misconceived," *Psychol. Bull.*, vol. 96, no. 2, pp. 394–401, 1984.
[108] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.
[109] J. Urbano, H. Lima, and A. Hanjalic, "A new perspective on score standardization," in *Proc. 42nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer, Eds. New York, NY, USA: ACM Press, 2019, pp. 1061–1064.
[110] C. J. van Rijsbergen, "Foundations of evaluation," *J. Document.*, vol. 30, no. 4, pp. 365–373, 1974.
[111] C. J. van Rijsbergen, *Information Retrieval*, 2nd Ed. London, U.K.: Butterworths, 1979.
[112] C. J. van Rijsbergen, "Retrieval effectiveness," in *Information Retrieval Experiment*, K. S. Jones, Ed. London, U.K.: Butterworths, 1981, pp. 32–43.
[113] P. F. Velleman and L. Wilkinson, "Nominal, ordinal, interval, and ratio typologies are misleading," *Amer. Statistician*, vol. 47, no. 1, pp. 65–72, Feb. 1993.
[114] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1998, pp. 315–323.
[115] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Inf. Process. Manage.*, vol. 36, no. 5, pp. 697–716, Sep. 2000.
[116] E. M. Voorhees, "Overview of the TREC 2004 robust track," in *Proc. 13th Text Retr. Conf. (TREC)*, E. M. Voorhees L. P. Buckland, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 2004, pp. 1–17.
[117] E. M. Voorhees, "Overview of the TREC 2005 robust retrieval track," in *Proc. 14th Text Retr. Conf. (TREC)*, E. M. Voorhees and L. P. Buckland, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 2005.
[118] E. M. Voorhees, "On building fair and reusable test collections using bandit techniques," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 407–416.
[119] E. M. Voorhees and D. K. Harman, "Overview of the seventh text REtrieval conference (TREC-7)," in *Proc. 7th Text Retr. Conf. (TREC)*, E. M. Voorhees and D. K. Harman, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 1998, pp. 1–24.
[120] E. M. Voorhees and D. K. Harman, "Overview of the Eigth text REtrieval conference (TREC-8)," in *Proc. 8th Text Retr. Conf. (TREC)*, E. M. Voorhees and D. K. Harman, Eds. Washington, DC, USA: National Institute of Standards and Technology (NIST), 1999, pp. 1–24.
[121] W. B. Ware and J. Benson, "Appropriate statistics and measurement scales," *Sci. Educ.*, vol. 59, no. 4, pp. 575–582, Oct. 1975.
[122] W. Webber, A. Moffat, and J. Zobel, "Score standardization for inter-collection comparison of retrieval systems," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 51–58.
[123] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
[124] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2006, pp. 102–111.
[125] E. Yilmaz, J. A. Aslam, and S. Robertson, "A new rank correlation coefficient for information retrieval," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 587–594.
[126] P. S. Yu, V. Tsotras, E. A. Fox, and C.-B. Liu, Eds., in *Proc. 15th Int. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM Press, 2006.
[127] F. Zhang, Y. Liu, X. Li, M. Zhang, Y. Xu, and S. Ma, "Evaluating web search with a bejeweled player model," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 425–434.
[128] F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma, "Models versus satisfaction: Towards a better understanding of evaluation metrics," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 379–388.

MARCO FERRANTE received the Ph.D. degree in functional analysis and applications from the Scuola Internazionale Superiore di Studi Avanzati (SISSA), in 1993. He is currently a Professor of probability theory at the University of Padua. His research interests include probability theory, stochastic calculus, stochastic processes, and their applications to biological models and information retrieval.

NICOLA FERRO (Senior Member, IEEE) is currently a Full Professor in computer science at the University of Padua. He is also a coordinator of the CLEF evaluation initiative. His research interests include information retrieval and its experimental evaluation, multilingual information access, and digital libraries.

NORBERT FUHR is currently a Full Professor in computer science at the University of Duisburg–Essen. His research interests include information retrieval, especially logic and probabilistic models, interactive retrieval, and evaluation methods.