ERR is not C/W/L: Exploring the Relationship Between Expected Reciprocal Rank and Other Metrics

Leif Azzopardi
University of Strathclyde
Glasgow, UK
leifos@acm.org

Joel Mackenzie
The University of Melbourne
Melbourne, Australia
joel.mackenzie@unimelb.edu.au

Alistair Moffat
The University of Melbourne
Melbourne, Australia
ammoffat@unimelb.edu.au

ABSTRACT

We explore the relationship between expected reciprocal rank (ERR) and the metrics that are available under the C/W/L framework. On the surface, it appears that the user browsing model associated with ERR can be directly injected into a C/W/L arrangement, to produce system measurements equivalent to those generated from ERR. That assumption is now known to be invalid, and demonstration of the impossibility of ERR being described via C/W/L choices forms the first part of our work. Given that ERR cannot be accommodated within the C/W/L framework, we then explore the extent to which practical use of ERR correlates with metrics that do fit within the C/W/L user browsing model. In this part of the investigation we present a range of shallow-evaluation C/W/L variants that have very high correlation with ERR when compared in experiments involving a large number of TREC runs. That is, while ERR itself is not a C/W/L metric, there are other weighted-precision computations that fit with the user model assumed by C/W/L, and yield system comparisons almost indistinguishable from those generated via the use of ERR.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; **Retrieval effectiveness**; **Presentation of retrieval results**; Relevance assessment.

KEYWORDS

Evaluation; effectiveness metric; user browsing model

ACM Reference Format:
Leif Azzopardi, Joel Mackenzie, and Alistair Moffat. 2021. ERR is not C/W/L: Exploring the Relationship Between Expected Reciprocal Rank and Other Metrics. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3471158.3472239

1 INTRODUCTION

Over the years numerous *information retrieval* (IR) evaluation metrics have been proposed, the goal being to measure the effectiveness of IR system result rankings [20]. During this time many commonalities between the metrics have been identified, with the majority calculating some variant of *expected utility* (also referred to as *expected rate of gain* in some of the associated literature) via a weighted-precision based measurement derived from a *user browsing model* [5, 17]. Most recently, the C/W/L framework has been proposed as a theoretically principled approach to describing and designing IR metrics, with new and existing metrics able to be encoded by defining a function describing the user population's *conditional continuation probability* [18, 19]. Under the C/W/L framework (pronounced "cool") it is possible to describe a range of both traditional and more novel metrics, including precision (P@k); average precision (AP); reciprocal rank (RR); discounted cumulative gain (DCG) [13]; rank-biased precision (RBP) [17]; time-biased gain (TBG) [22]; INSQ and INST [18, 19]; bejeweled player model (BPM) [28]; information foraging theory (IFT) [1]; and data driven metrics (DDM) [3].

The C/W/L framework enables the measurement of expected utility (as noted, also referred to as the expected rate of gain) and has a number of advantages:

(i) measurements are in defined units of "expected gain accrued per document inspected" and can be compared between metrics, meaning that (for example) a metric score computed via RBP is directly comparable to the scores from TBG, BPM, INST, IFT, and so on;
(ii) the C/W/L framework enables computation of a range of further values, including *expected total utility* (referred to in some of the literature as *expected total gain*), the *expected total cost*, and the *expected search depth*;
(iii) some of those attributes of the user's modeled activity when perusing the document ranking can be compared directly to observational studies (for example, *items viewed* [25, 29], or *time spent* and *last item examined* [1, 3]), allowing parameters to be fitted to different types of search behavior, and hence more specific measurement to be undertaken;
(iv) new metrics can be encoded by instantiating an appropriate user model, and then specifying how to compute the conditional continuation probability function implied by the model; and
(v) there is a natural extension that incorporates search sessions, in which multiple queries are issued and the user's overall goal is used as part of the user browsing model [26, 27].

When taken together, these five attributes mean the C/W/L framework provides an extensible and versatile basis for measuring different aspects of retrieval performance under different user modeling assumptions, with the flexibility provided by point (iv) perhaps of greatest importance.

While the C/W/L framework is both appealing and flexible, it is by no means universal, and a range of other models have emerged

ICTIR '21, July 11, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8611-1/21/07…$15.00
https://doi.org/10.1145/3471158.3472239

[11, 12]. One popular metric often used when measuring web search effectiveness – *expected reciprocal rank* (ERR) – offers a fundamentally different user browsing model and measurement of a ranking, computing the expected inverse of the effort to find a single relevant item [6]. Initially, it was thought that ERR could be seamlessly fitted into the C/W/L framework (for example, see the claim made by Moffat et al. [19, top of page 15]). We carefully consider that question in the current work, and as a first result, provide a demonstration that in fact ERR cannot be described as a C/W/L metric.

Given that context, our second objective is to explore the extent to which it is possible to construct a C/W/L-compliant and ERR-inspired metric that shares its evaluation properties with ERR, but also inherits all of the other C/W/L benefits listed above. As we show below, we can describe C/W/L metrics that closely mirror the retrieval evaluations and system comparisons attained by ERR.

Note that we do not in any way suggest that ERR is redundant. Its user model is simply based on different user browsing assumptions, and hence it provides an alternative to C/W/L-based metrics. But in terms of practical use, it appears that the impact of the different browsing model can be minimized, a point that might be of interest to theoreticians and practitioners alike.

2 ERR AND OTHER METRICS

Information retrieval has a long tradition of practical experimentation, see, for example, Sanderson [20] for an overview. A key part of that endeavor has been the development of *effectiveness metrics*, computations that take an ordered sequence of documents as returned by a retrieval system (a *run*), and a set of determinations as to the relevance of each of the possible documents that might appear in the run (a set of *qrels*), and combine them to generate various numeric scores that summarize the merits of that run. For any given system, the run scores can then be aggregated over a fixed set of queries (or *topics*); and sets of systems can then be compared by mean score; or via the computation of a risk-adjusted mean score [4, 8, 10]; or via application of an appropriate statistical test; or by a combination of all three.

User Browsing Models. The last decade has seen several proposals for new metrics. One that has received a lot of use is *expected reciprocal rank* (ERR) [6]. It was developed as a generalization of the earlier *reciprocal rank* (RR) metric, in response to the emergence of *graded relevance judgments* [13]; they, in turn, were a generalization of the previous predominantly binary judgments. Expected reciprocal rank is intended to reflect the way that users approach ranked document lists, and was one of two proposals that pioneered an explicit focus on user browsing models, the second being the *rank-biased precision* (RBP) metric of Moffat and Zobel [17].

In both RBP and ERR each user is presumed to commence at the first document in the ranking, and after considering it in some way, to either proceed to the second with some defined probability, or to exit the run after only looking at the first document. More generally, the fraction of users that reach the i th document in the ranking are presumed to either abandon the run at that point, or to continue to the $i+1$ th document with some calculable probability. These probabilistic sequential browsing models of user behavior are also sometimes called *cascade models* [9].

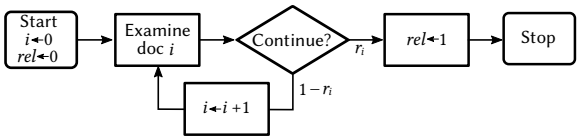

Figure 1: The user browsing model associated with ERR.

Chapelle et al. [6] define ERR as:

$$M_{ERR}(\mathbf{r}) = \sum_{i=1}^{\infty} \left(\frac{r_i}{i} \cdot \prod_{j=1}^{i-1} (1 - r_j) \right), \tag{1}$$

where $\mathbf{r} = \langle r_i \rangle$ is derived from the relevance categories with which the documents in the run have been labeled (the qrels). For example, one common way of converting categorical relevance labels g_i across G distinct classes (for example, $g_i \in \{0 \ldots (G-1)\}$ with $G = 4$ and the four categories representing "not at all relevant", "somewhat relevant", "relevant", and "highly relevant" respectively) to numeric gain values r_i is via the gain mapping:

$$r_i = \frac{2^{g_i} - 1}{2^{G-1}}. \tag{2}$$

With this gain mapping, even maximally relevant documents ($g_i = 3$ when $G = 4$) are treated as being sometimes not recognized as such by the individual searchers, with the probability of that happening set at $1 - 7/8 = 0.125$; conversely, "somewhat" relevant documents are regarded as being sufficiently useful that they satisfy some of the individual searchers, also with probability $1/8 = 0.125$. This particular gain mapping also stipulates that users never deem any of the "not at all relevant" documents to be acceptable.

In terms of modeled behavior, ERR supposes that each user ends their inspection of the run as soon as they encounter a document that they believe is relevant, with r_i being the probability that the user will regard the i th document in the ranking as fully satisfying their information requirement. This structure in illustrated in Figure 1. Because the gain mapping depicted by Equation 2 never allows $r_i = 1$, a non-zero fraction of the user population is regarded as examining the run indefinitely, which is why the summation in Equation 1 must, in theory at least, run to infinity. In practice, the summation is often truncated to some depth k. For example, ERR@20 evaluates the summation across the first $k = 20$ terms, and ignores any contribution from the remaining tail of the distribution.

More precisely, Equation 1 computes a weighted sum, over a population of users and across all depths in the ranking, of the probability of exiting at rank i with one unit of perceived gain, in each case divided by the number of elements from the ranking inspected by that user. That is, ERR is defined as the expected effort in order to obtain one unit of (subjective) gain [6], and hence, like the C/W/L family, ERR has units of "expected gain accrued per document inspected". The total gain that each user derives from their search – or rather, their perceived (or deemed) gain – is always exactly one unit of relevance. Finally, note that RR can be viewed as the restriction of ERR to binary qrels – Equation 1 still exactly describes the computation, but for RR there are only two possible values, and r_i must always be one of $\{0, 1\}$.

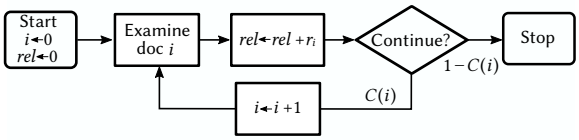

Figure 2: The user browsing model associated with C/W/L metrics. Adapted from similar diagrams by Moffat et al. [17, 18].

The C/W/L Framework. As noted above, one important way of interpreting effectiveness metrics is through the lens of the corresponding user browsing model, which represents how each individual user, and hence a whole population of users, interacts with the run [5, 6, 14, 17, 18, 28]. Moffat et al. [18, 19] formalized that idea into a framework they call "C/W/L", which provides a mathematical foundation for effectiveness measurement. The next few paragraphs summarize the main elements of their proposal.

Members of the C/W/L family are most intuitively defined via a function $C(i)$ (there are two others, $W(i)$ and $L(i)$, hence the name; with the other two determined once any one of the three is defined) which specifies the conditional probability of a user continuing from the document at rank i to the document at rank $i+1$. The user's propensity to continue may be based on any/all of i, the depth in the ranking; r_i, the gain associated with the i th document in the ranking; the total gain accrued through until rank i; and the user's initial search goal, measured in units of "relevance" [18, 19]. Other factors known to the user at the time that they make that i th decision to continue or stop can also be included if appropriate.

Figure 2 describes this user browsing model. Note the differences between it and the ERR model shown in Figure 1: in C/W/L metrics, gain can be accumulated from every document observed, rather than from only the last document observed; and r_i is the fractional gain accrued by viewing the i th document, rather than the probability of that document being regarded as fully relevant.

The actual value of a C/W/L metric is computed as the inner product of two vectors:

- the vector of per-document gains, $\mathbf{r} = \langle r_i \rangle$; and
- a vector of weights, $\langle W(i) \rangle$, where $W(i)$ is the fraction of all user attention that is paid to the document at rank i.

That is, the C/W/L metric value is computed as

$$M(\mathbf{r}) = \sum_{i=1}^{\infty} W(i) \cdot r_i \,. \tag{3}$$

The value of $W(i)$ is derived from $C(i)$ via an intermediate function $V(i)$, which records the fraction of the population of users that view the document at rank i:

$$V(i) = \begin{cases} 1 & \text{if } i = 1 \\ C(i-1) \cdot V(i-1) & \text{otherwise} \,. \end{cases}$$

Given the vector $\langle V(i) \rangle$, the value of

$$V^+ = \sum_{i=1}^{\infty} V(i)$$

is computed, the sum over all ranks i of the proportion of users that view that i th document; and then

$$W(i) = \frac{V(i)}{V^+}$$

is used to generate the weight vector $\langle W(i) \rangle$. That is,

$$M(\mathbf{r}) = \sum_{i=1}^{\infty} \frac{V(i) \cdot r_i}{V^+} \,. \tag{4}$$

It is important to note that if $\sum_{i=1}^{\infty} r_i$ is finite (that is, if there is a finite number of documents judged relevant or partially relevant) and V^+ is infinite (that is, the function $C(i)$ is such that the sequence of partial sums $\sum_{j=1}^{i} V(j)$ does not approach a limiting value as i increases) then the metric value will approach zero. In such a situation the user can only gain a finite total amount of relevance, but is modeled (with some non-zero probability) as continuing to examine documents indefinitely, making their expected utility (measured in the desired units of "gain accrued per document inspected") approach zero.

As an example of a C/W/L metric, consider reciprocal rank (RR) with binary qrels, that is, with $r_i \in \{0, 1\}$. It can be C/W/L-defined via the conditional continuation function

$$C_{\text{E5}}(i) = (1 - r_i) \,. \tag{5}$$

Here the user is modeled as always proceeding to the next document if $r_i = 0$, and always stopping if that i th document is relevant. Suppose that d is the shallowest rank at which $r_d = 1$. Then $V(i) = 1$ for $1 \le i \le d$ and $V(i) = 0$ thereafter; and hence $V^+ = d$, and $W(i) = 1/d$ for $1 \le i \le d$. The ERG (expected rate of gain) metric $M(\mathbf{r})$ associated with the ranking is then $1/d$. Or, if there is no d in the ranking for which $r_d = 1$, the user is modeled as continuing endlessly, and never exiting the ranking; the corresponding metric score is zero. Hence, since RR can be completely described by a $C(\cdot)$ function, it is a member of the C/W/L family.

ERR is Not C/W/L. Given the strong relationship between RR and ERR, it is tempting to conclude that ERR must also be a C/W/L metric, defined in exactly the same way, that is, via Equation 5. Indeed, Moffat et al. [19] made exactly that claim, and only later recognized its incorrectness[1].

To demonstrate that ERR is *not* a C/W/L metric, consider the gain vector $\mathbf{r} = \langle \alpha, \alpha, \alpha, \ldots \rangle$, that is, every document in the run provides exactly the same gain of $0 < \alpha < 1$. Because the C/W/L weights $W(i)$ must sum to 1.0, it is clear that *every* C/W/L metric must give a score of α to the run described by $\mathbf{r}$. No matter what behavior the users exhibit, their expected rate of gain is always exactly α per document inspected.

Now consider the value that arises for ERR (Equation 1). The summation becomes

$$\sum_{i=1}^{\infty} \left(\frac{\alpha}{i} \prod_{j=1}^{i-1} (1-\alpha) \right) = \alpha + \sum_{i=2}^{\infty} \left(\frac{\alpha}{i} \prod_{j=1}^{i-1} (1-\alpha) \right) > \alpha \,,$$

with the final inequality holding because $0 < \alpha < 1$. That is, for this ranking there is *no* $C(\cdot)$ function which yields the same numeric

[1] See https://people.eng.unimelb.edu.au/ammoffat/abstracts/mbst17acmtois-errata.pdf, dated September 2019, and accessed May 2021 while preparing this work. That note provides the argument that is presented here.

score as Equation 1, and hence, by existence of a counter-example, ERR is not a C/W/L metric.

3 SHALLOW ADAPTIVE C/W/L METRICS

Two immediate questions then arise:

(1) given that the function $C_{\text{E5}}(i) = (1 - r_i)$ (Equation 5) does not correspond to ERR, then what are the properties of the metric that it *does* instantiate; and
(2) does Equation 5, or some other $C(i)$ function, provide a useful C/W/L approximation of ERR, where "useful" and "approximation" need specification.

In this section we explore five ways in which an ERR-like metric might be formulated within the C/W/L framework. The properties of these five options are then compared in Section 4.

Option One. The first possibility for an "ERR-inspired" C/W/L variant is to simply take the $C_{\text{E5}}(\cdot)$ function defined in Equation 5, and use it to specify a metric.

However $C_{\text{E5}}(\cdot)$ has a problem: the sum V^+ does not have a finite limit. In particular, if there is a tail in the ranking beyond some depth d for which $V(d) > 0$ and $r_j = 0$ for $j \geq d$, then $V^+ = \sum_{i=1}^{\infty} V(i) = \infty$ is unbounded, because some non-zero fraction of the users are modeled as examining every document in the collection. Those users continue on patiently, viewing document after document, depressing the "expected gain accrued per document inspected" towards zero.

Moreover, the conditions required for this scenario (namely, that $V(d) > 0$ and $r_j = 0$ for $j \geq d$ at some point d in the run) are commonplace when graded judgments are being used. For example, Section 2 noted (Equation 2) that one common way of converting categorical relevance labels g_i to gain values r_i has $r_i \leq 7/8$ and hence $C(i) \geq 1 - (7/8) = 0.125$ at every depth i. That, in turn, means that $V(d) > 0$ at the point d in the ranking of the last judged document. In other words, regardless of the judged documents prior to depth d, Equation 5 cannot guarantee bounded values for V^+, and hence all derived metric scores will tend to zero when computed to depth infinity (Equation 3).

Option Two. It is also tempting to consider that:

$$C_{\text{E6}}(i) = \frac{i}{i+1}(1 - r_i) \tag{6}$$

might provide a useful approximation to ERR, with $V(i)$ explicitly equated to the "effective discount" part of ERR [6] and incorporated into the $C(\cdot)$ and hence $V(\cdot)$ functions via a telescoping product:

$$V(i) = \frac{1}{i} \prod_{j=1}^{i-1} (1 - r_j)\,. \tag{7}$$

Unfortunately, this option is not suitable either, as the relationship:

$$\sum_{i=1}^{k} \frac{1}{i} \approx \ln k + 0.577$$

is well-known[2], and means that the derived value V^+ corresponding to Equation 7 is also non-convergent. That is, even though it grows much more slowly than was the case with Equation 5, Equation 6 can also not be used as the basis for a (well-defined) C/W/L metric.

[2]See, for example, https://en.wikipedia.org/wiki/Euler-Mascheroni_constant.

Indeed, Equations 5 and 6 yield values for V^+ that are bounded if and only if:

- the gain mapping in use allows $r_i = 1$ when $g_i = G - 1$; and
- there is a document in the ranking for which $g_i = G - 1$, and thus $r_i = 1$.

In all other cases the C/W/L metric scores, computed to rank ∞ (Equation 4) must have zero as their limiting values, regardless of the gain values of the documents prior to that limiting depth d that marks the point beyond which no judgments are available.

Finally, note that non-convergence issues don't arise with reciprocal rank (RR) and binary relevance judgments: the binary gain mapping *does* allow $r_i = 1$, and if there is no such document in the ranking, the correct metric score is simply deemed to be zero. It is the use of graded relevance judgments that means that ERR cannot be a C/W/L metric.

Truncation at k. As is already done with other non-convergent metrics – notably precision@k and DCG@k – one possible remediation is to evaluate the metric to a fixed depth, so that V^+ can be bounded even when the ranking has $r_i = 0$ for all depths i. For example, Equation 5 can be readily modified so that no users proceed beyond the k th document in the ranking:

$$C_{\text{E8}}(i) = \begin{cases} (1 - r_i) & \text{when } i < k \\ 0 & \text{when } i \geq k\,. \end{cases} \tag{8}$$

The corresponding browsing model has users "giving up" at depth k, and is somewhat ad hoc, with k becoming a parameter that governs the metric score and imposes a hard limit on evaluation depth. But it is also true that pooled judgments are often generated to some finite depth, and so evaluating rankings to that same depth might be deemed to be a reasonable compromise.

Similarly, a cutoff can also be applied to Equation 6:

$$C_{\text{E9}}(i) = \begin{cases} i \cdot (1 - r_i)/(i+1) & \text{when } i < k \\ 0 & \text{when } i \geq k\,. \end{cases} \tag{9}$$

Option Three. Given that Equations 5 and 6 have limitations and must be transformed into Equations 8 and 9 to be useful in the C/W/L framework, are there any "infinite depth" ERR-inspired formulations for $C(i)$ that *can* be considered?

That question is answered in the affirmative if compound functions are used, and a multiplicative term is introduced that guarantees that V^+ is finite. For example, it is possible to combine an RBP-like (employing a parameter ϕ) continuation function with Equation 5, to obtain:

$$C_{\text{E10}}(i) = \phi \cdot (1 - r_i) \tag{10}$$

where $0 \leq \phi < 1$ forces convergence in the same way as it does in RBP itself, meaning that arbitrary truncation at depth k is not required.

Option Four. Similarly, inspiration can be taken from INSQ [18], defining

$$C_{\text{E11}}(i) = \left(\frac{i + 2T - 1}{i + 2T}\right)^2 \cdot (1 - r_i)\,, \tag{11}$$

where $T \geq 0$ in conjunction with the squaring of the fraction forces convergence, and ensures that V^+ is bounded even when the summation is taken to ∞. (In INSQ, T is the total "volume of relevance" the user is hoping to acquire as a result of their search.)

Option Zero. Finally, we also add into the set of possibilities (as a kind of base case, hence the terminology "option zero") the function $C_{\text{RBP}}(i) = \phi$, the definition of rank-biased precision [17]. It is not in any way inspired by ERR, but when ϕ is small (say, $\phi \approx 0.5$), RBP is also a heavily top-weighted effectiveness metric that can employ graded relevance via a gain mapping, and hence provides a useful reference point against ERR.

4 EXPERIMENTS

Given the different ERR-inspired C/W/L alternatives, we now explore the second question presented in Section 3. First, we aim to determine whether any of the possible options produce a measurement that approximates the ERR measurement (that is, yields comparable metric values in a numeric sense); and then second, determine their usefulness, in terms of whether they lead to the same or similar system orderings as ERR (that is, comparable system orderings in an experimental outcomes sense).

Experimental Setup. To perform the analysis we used the Ad-Hoc runs submitted to the 2010 and 2011 TREC Web Tracks [7, 23], both of which used ERR@20 as the official metric. The 2010 and 2011 tracks contain 56 and 38 runs respectively for each of 50 topics in each year. We included both runs utilizing the whole ClueWeb09 corpus and those making use of only the ClueWeb09B subset.

The qrels for these two Web Tracks have five levels of relevance: grade −2 denoting spam; grade 0 denoting non-relevant; and grades 1, 2, and 3 representing increasing levels of relevance. In our experiments we mapped grade −2 back to 0, and then used a four-category relevance scale ($G = 4$) to calculate gain, employing the mapping presented in Equation 2. Expected reciprocal rank scores were then computed using the official `gdeval` tool[3]; with all of the C/W/L metrics computed within the `cwl_eval` framework [2].

Figure 3 provides an overview of the resulting ERR@20 scores. The gain mapping results in a reasonable spread of scores, but with more ERR@20 below 0.25 than above it, and with a maximum (for this gain mapping) possible ERR (and ERR@20) score of 0.9347.

To ensure that our analysis was not influenced by unjudged documents, we next computed residuals (see Moffat et al. [17, 19]) across all system/topic pairs submitted to the 2010 Track, first calculating an ERR@20 score in the usual manner (taking unjudged documents as relevance grade 0), then calculating full-depth ERR taking unjudged documents to have a grade of 3, and then taking the difference between those two scores. All system/topic combinations with a residual > 0.05 were discarded, resulting in the retention of 1,760 system/topic pairs (around 65% of the original set of runs).

Approximating ERR@20. To evaluate the parity between measures, we used the 2010 Web Track runs, holding out the 2011 runs for the system comparison.

To determine how closely the C/W/L metrics approximate ERR, we computed the correlation between the measurements taken for each option and ERR, using both Pearson's r and Spearman's ρ. Each of the C/W/L metrics has a parameter, and an exhaustive search was performed to find the parameter value that maximized the correlation coefficient between the measured ERR@20 scores and corresponding C/W/L scores for each of the five shallow options.

[3] https://trec.nist.gov/data/web/10/gdeval.pl

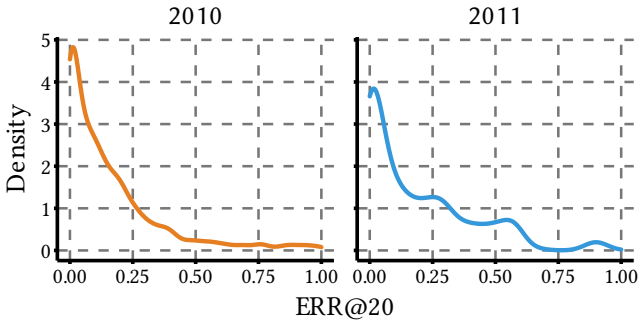

Figure 3: Observed ERR@20 score distributions for the TREC 2010 and TREC 2011 submitted runs, calculated on a per-system and per-topic basis. A total of 2,688 runs (system-topic combinations) are included for TREC 2010, and 1,900 for TREC 2011.

Table 1: Various metrics and their settings to achieve a maximal Pearson's r or Spearman's ρ with respect to ERR@20, using exhaustive search and the TREC 2010 runs. All correlations are highly significant, with $p < 0.01$.

Metric	Param.	Value	Pearson's r	Value	Spearman's ρ
RBP	ϕ	0.50	0.966	0.60	0.990
C_{E8}	k	3	0.936	5	0.944
C_{E9}	k	7	0.955	20	0.999
C_{E10}	ϕ	0.62	0.955	0.70	0.993
C_{E11}	T	1.25	0.953	1.35	0.995

Table 1 reports the correlation coefficients and the corresponding maximizing parameter values, and confirms that very high score-based correlations can be achieved. Figure 4 plots those same relationships, with curvilinear relationships emerging between ERR@20 (which cannot exceed 0.9347) and the various C/W/L inspired variants (all of which are bounded above by 0.875 with this gain mapping). Figure 5 shows the correlation robustness as the parameter ϕ for RBP and C_{E10} (Equation 10) is varied. The maximizing values shown in Table 1 are at the highest point of each of the respective curves, but there is a broad band of parameter that yields high correlations, and the exact choice of ϕ within that broad band is relatively unimportant.

These results illustrate very clearly that all of the five ERR-inspired C/W/L metrics, including RBP, can be configured to correlate highly with ERR@20, with the resulting parameters, as expected, biased towards shallow evaluation. That is, from these experiments we can conclude that ERR can be numerically closely approximated by C/W/L-structured shallow metrics.

System Orderings. To determine how well the proposed metrics correlate with ERR@20 in terms of ordering *systems*, we switch to the (as yet unused) TREC 2011 Web Track runs, consisting of 38 unique runs over 50 topics. We first scored each run with ERR@20 to define the *ground-truth* system ordering, and then repeated this process using each of the five proposed C/W/L metrics, taking the parameters for each metric that were established using the TREC

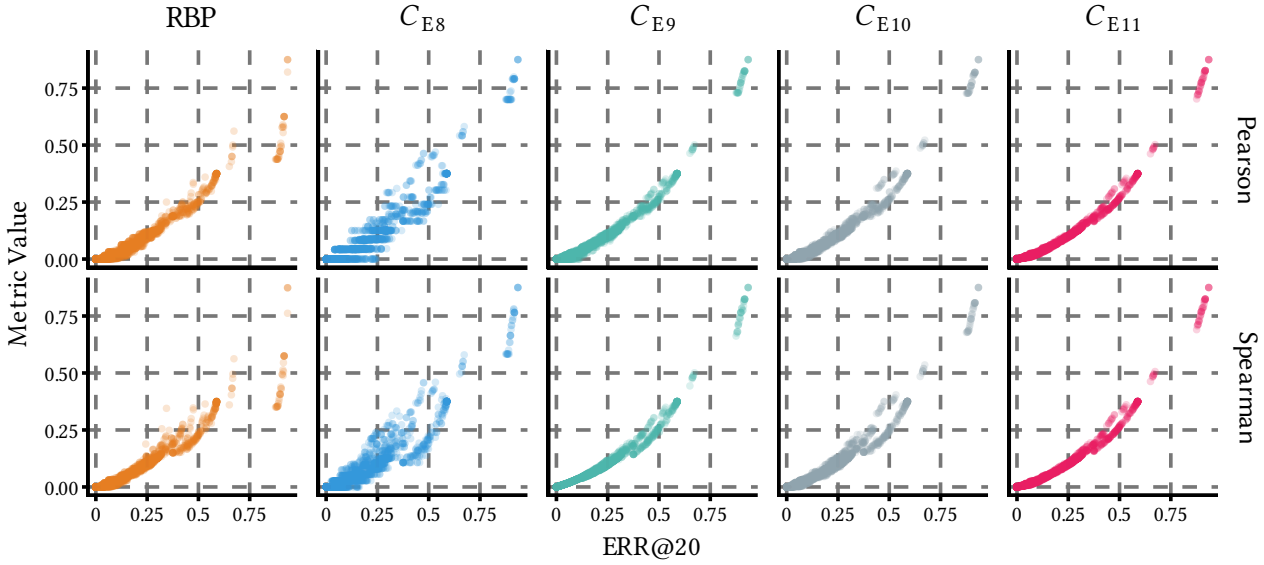

Figure 4: Correlations between ERR@20 and five C/W/L metrics, with parameters set via minimization of Pearson's r (top) or Spearman's ρ (bottom). Each pane shows a total of 1,760 runs (system-topic combinations). All correlations are highly significant, with $p < 0.01$.

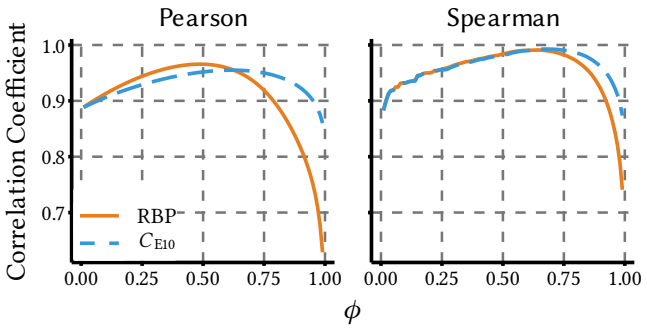

Figure 5: Sensitivity of two correlation coefficients between ERR@20 and RBP, and between ERR@20 and C_{E10} as the RBP/C_{E10} parameter ϕ is varied, using the filtered TREC 2010 Web Track runs. A total of 1,760 runs (system-topic combinations) are included, all with ERR@20 residuals less than 0.05.

2010 runs (for Spearman's ρ, see Table 1). We then computed the correlation between system orderings for all metric pairs using both a plain Kendall's τ [15], and a top-weighted Kendall's τ in which the weight of the system at rank s was set to $1/(s+1)$ (see Shieh [21] and Vigna [24]). Table 2 shows the results.

As can be seen in the table, the system orderings induced by the different metrics are almost all greater than 0.9, indicating close agreement between the metrics. Moreover, the top-weighted correlations in the lower half of the table tend to be higher than the corresponding values in the top half of the table, providing evidence that the six metrics agree closely on high-scoring systems, and that their differences tend to be in terms of the relative ordering of the low-scoring systems.

Table 2: Kendall's τ between different system orderings on TREC 2011 for all metric pairs using parameters derived from Spearman's ρ on TREC 2010. Entries above the diagonal represent unweighted correlations; entries below the diagonal represent weighted correlations. All correlations are highly significant, with $p < 0.01$.

	ERR@20	RBP	C_{E8}	C_{E9}	C_{E10}	C_{E11}
ERR@20	—	0.932	0.909	0.903	0.909	0.906
RBP	0.946	—	0.898	0.875	0.881	0.878
C_{E8}	0.940	0.949	—	0.909	0.926	0.906
C_{E9}	0.960	0.921	0.941	—	0.977	0.986
C_{E10}	0.944	0.941	0.967	0.972	—	0.980
C_{E11}	0.960	0.921	0.941	0.995	0.974	—

Retrospective Validation. Our final experiment was to retrospectively validate the parameter choices for each of the metrics. In particular, we repeat the earlier experiment where we maximize the correlation between the ERR@20 run scores, and the proposed metric scores, measured by either Pearson's r or Spearman's ρ, but now using the 2011 Web Track data, in order to provide a post-hoc confirmation of parameter stability (or not) between the two different datasets. The maximizing parameters for each of the metrics over the 2011 Web Track were indeed very similar to those found for the 2010 data (Table 1). For example, the maximizing parameter values for RBP measured with Pearson's r or Spearman's ρ were $\phi = 0.46$ and $\phi = 0.65$ respectively, very close to the values showing in Table 1. Similarly, for TREC 2011 the parameters found for C_{E11} were $T = 1.09$ and $T = 1.23$, differing by only a small amount from those shown in Table 1. Note that Table 2 was constructed before this final phase of experimentation was undertaken, and that Table 2 used the parameters developed using (only) the TREC 2010

data (Table 1), maintaining the clear separation between training data and test data.

5 SUMMARY AND FUTURE DIRECTIONS

We have contrasted the user browsing models, and hence properties, of a range of shallow effectiveness metrics, with an emphasis on ERR-like approaches, and on web search applications. While ERR cannot be described within the C/W/L framework, and its user browsing model is distinct from that of the C/W/L approach, it generates run scores that can be closely mirrored by a range of C/W/L-compliant effectiveness metrics, and generates system orderings that can likewise be closely matched. Moreover, we demonstrated that the parameters required to obtain that similar behavior are relatively stable, suggesting that the relationship between ERR and the shallow C/W/L metrics is a robust one.

Behind all of these comparisons is, of course, a more fundamental question – are either of the ERR user browsing model or the C/W/L user browsing model what users *actually* do, and if not, what other factors not already taken into account might influence their behavior? For example, the last document viewed may have a disproportionate influence on the overall user perception of the ranking [16], and focusing on it may be a way of bringing ERR-style metrics closer to the scores computable from the C/W/L gain accumulation approach.

New results – including careful measurements based on observations of users carrying out genuine search tasks, and selecting model parameters fitted against those observations – can be expected to continue to emerge as we build a better understanding of how users interact with search results rankings while they are carrying out their varied search tasks.

Acknowledgments. This work was supported by the Australian Research Council (project DP200103136).

REFERENCES

[1] L. Azzopardi, P. Thomas, and N. Craswell. Measuring the utility of search engine result pages: An information foraging based measure. In *Proc. SIGIR*, pages 605–614, 2018.

[2] L. Azzopardi, P. Thomas, and A. Moffat. cwl_eval: An evaluation tool for information retrieval. In *Proc. SIGIR*, pages 1321–1324, 2019.

[3] L. Azzopardi, R. W. White, P. Thomas, and N. Craswell. Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proc. CHIIR*, pages 213–222, 2020.

[4] R. Benham, B. Carterette, J. S. Culpepper, and A. Moffat. Bayesian inferential risk evaluation on multiple IR systems. In *Proc. SIGIR*, pages 339–348, 2020.

[5] B. Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proc. SIGIR*, pages 903–912, 2011.

[6] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. CIKM*, pages 621–630, 2009.

[7] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. In *Proc. TREC*, 2010.

[8] K. Collins-Thompson. Accounting for stability of retrieval algorithms using risk-reward curves. In *Proc. SIGIR*, pages 27–28, 2009.

[9] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. WSDM*, pages 87–94, 2008.

[10] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proc. SIGIR*, pages 23–32, 2014.

[11] M. Ferrante and N. Ferro. Exploiting stopping time to evaluate accumulated relevance. In *Proc. ICTIR*, pages 169–176, 2020.

[12] M. Ferrante, N. Ferro, and M. Maistro. Towards a formal framework for utility-oriented measurements of retrieval effectiveness. In *Proc. ICTIR*, pages 21–30, 2015.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20:2002, 2002.

[14] J. Jiang and J. Allan. Adaptive effort for search evaluation metrics. In *Proc. ECIR*, pages 187–199, 2016.

[15] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[16] J. Liu and F. Han. Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. In *Proc. SIGIR*, pages 1141–1150, 2020.

[17] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2:1–2:27, 2008.

[18] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[19] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.

[20] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. & Trends in IR*, 4(4):247–375, 2010.

[21] G. S. Shieh. A weighted Kendall's tau statistic. *Statistics & Probability Letters*, 39 (1):17–24, 1998.

[22] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[23] I. Soboroff, N. Craswell, C. L. A. Clarke, and G. V. Cormack. Overview of the TREC 2011 Web Track. In *Proc. TREC*, 2011.

[24] S. Vigna. A weighted correlation index for rankings with ties. In *Proc. WWW*, pages 1166–1176, 2015.

[25] A. F. Wicaksono and A. Moffat. Empirical evidence for search effectiveness models. In *Proc. CIKM*, pages 1571–1574, 2018.

[26] A. F. Wicaksono and A. Moffat. Metrics, user models, and satisfaction. In *Proc. WSDM*, pages 654–662, 2020.

[27] A. F. Wicaksono and A. Moffat. Modeling search and session effectiveness. *Inf. Proc. & Man.*, 58(4):102601, 2021.

[28] F. Zhang, Y. Liu, X. Li, M. Zhang, Y. Xu, and S. Ma. Evaluating web search with a bejeweled player model. In *Proc. SIGIR*, pages 425–434, 2017.

[29] Y. Zhang, L. A. F. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Inf. Retr.*, 13(1):46–69, 2010.