# On The Pluses and Minuses of Risk

Rodger Benham[1], Alistair Moffat[2], and J. Shane Culpepper[1]

[1] RMIT University, Melbourne, Australia
[2] The University of Melbourne, Melbourne, Australia

**Abstract.** Evaluating the effectiveness of retrieval models has been a mainstay in the IR community since its inception. Generally speaking, the goal is to provide a rigorous framework to compare the quality of two or more models, and determine which of them is the "better". However, defining "better" or "best" in this context is not a simple task. Computing the average effectiveness over many queries is the most common approach used in Cranfield-style evaluations. But averages can hide subtle trade-offs in retrieval models – a percentage of the queries may well perform worse than a previous iteration of the model as a result of an optimization to improve some other subset. A growing body of work referred to as *risk-sensitive evaluation*, seeks to incorporate these effects. We scrutinize current approaches to risk-sensitive evaluation, and consider how risk and reward might be recast to better account for human expectations of result quality on a query by query basis.

## 1 Introduction

Risk measures have emerged in IR in response to the goal of improving a system without negatively impacting the user's experience of the system's overall effectiveness. This is an issue because measured effectiveness is usually volatile across both systems and topics. That is, selecting one system over another because it has a higher mean effectiveness could be *risky*, as the mean may well disguise substantial variability of the system effectiveness across the range of queries. Several approaches have been proposed to quantitatively measure the tension between risk and reward: URisk [6, 24], TRisk [10], ZRisk and GeoRisk [9].

Common to all of these is that they measure risk-reward trade-offs piecewise, with effectiveness decreases penalized by a linear factor, and hence with the "loss" rate for small erosions in effectiveness the same as for the rate for wholesale decreases. Previous studies have demonstrated that users are unable to discern small changes in effectiveness scores, so an interesting question is whether small losses should count as much as large losses. For example, Allan et al. [2] observe that bpref effectiveness and recall follow an S-shaped pattern, where there is a "large intermediary region in which the utility difference is not significant". Similar effects have been observed in the field of economics, and the application of an S-shaped weighting function for modeling the psychological value of monetary gains and losses has been proposed by Tversky and Kahneman [20], with losses perceived as being twice as costly in a negative sense as similar-sized gains are in

a positive sense. In IR-related experimentation, Turpin and Scholer [19] found that the only reliable signal of whether retrieval effectiveness scores impacted task performance was precision at depth one.

Here we explore whether an S-shaped risk function that strongly weights outliers produces different system orderings than the linear risk function that is embedded in all current proposals. If that were the case, running a user study to ascertain the shape of the applicable trade-off function for the IR domain would be an important next step in improving risk measures. We also take a broader view of the meaning of "risk", and in doing so, conclude that current terminology is potentially ambiguous and can be improved. In response, we propose changes to how practitioners discuss risk-based trade-offs, and further suggest reversing the sign of risk-inclusive evaluation results when they are reported.

## 2 Background

In an investment portfolio, risk (sometimes known as "beta") is compared against expected gain (referred to as "alpha") to distinguish between investment options that are safe and reliable but low return, and more speculative options that are potentially high-return, but also have a higher probability of leading to losses. Risk can be spread in this context, with a portfolio as a whole being acceptable if it makes the expected level of return, even if some components within it perform poorly. In IR, however, a user may abandon a search service that returns rankings of variable quality [22], even if its overall "mean" behavior is better than that of its competitors. Collins-Thompson [8] first demonstrated the utility of incorporating risk measures used by economists in IR evaluation, borrowing from the practice of forming risk-reward curves. In experiments in which the risk function counted the number of relevant documents lost due to query expansion failure, Collins-Thompson showed that two systems with the same mean effectiveness might possess "very different risk profiles."

Wang and Zhuhan [23] used a mean-variance approach to perform risk-sensitive retrieval, by modifying the language modeling formula to accept a parameter $b$ to indicate the risk preference of the user, with the document selection problem modeled similarly to the investment selection technique of portfolio theory. The key idea is that if the documents at the head of the SERP are similar, and one is not relevant, then they all might be poor choices. Risk is then spread by diversifying the elements at the head of the result list. Although Wang and Zhuhan did not define a measure that could be used to instrument risk-reward profiles, their approach is an example of how lessons learned in economics might be applied in a retrieval model to accomplish a similar goal.

Wang et al. [24] proposed an approach for quantitatively measuring risk based on the differences in scores. The approach was later used in the TREC 2013 and 2014 Web Tracks under the alias URisk [6, 7], which was adopted more readily than the original name, $T_\alpha$. To calculate URisk, an experimental system is compared against a baseline system using the formula:

$$URisk_\alpha = (1/c) \cdot \left[ \sum Wins - (1+\alpha) \cdot \sum Losses \right], \tag{1}$$

where a "win" is a case where the difference in score is positive for the experimental system, and a "loss" is the reverse, and where $c$ is the number of paired comparisons. The parameter $\alpha$ is user-selected, and linearly scales the relative impact of losses, so that the computed value is an adjusted mean difference. Positive URisk values indicate that the experimental system comes out ahead on balance, conversely, negative values are indicative of risk. The URisk formula can be used as a cost function in learning-to-rank [24].

The TREC evaluation exercises demonstrated the practical applicability of using the URisk measure, which led to several alternative formulations of risk-sensitivity. An issue with URisk values was that although it was clear when an experimental system survived the risk threshold, it was unclear whether it was statistically significant. Dinçer et al. [10] proposed TRisk to solve this problem, which is a studentized version of URisk that can be used to perform an inferential risk and reward analysis between two systems, defined as:

$$TRisk_\alpha = URisk_\alpha / SE(URisk_\alpha), \tag{2}$$

where *SE* is the standard error of the URisk sampling distribution. Like URisk, TRisk compares an experimental system against a baseline system, but computes a $t$-value which incorporates both mean and variance. When $t < -2.0$ (two standard errors), changing to the experimental system would give rise to significant risk, and when $t > 2.0$, a change to the experimental system would allow a significant reward.

Rather than a pair of systems, empirical studies often compare multiple systems. For example, Zhang et al. [28] propose a graphical evaluation approach to assess the bias-variance relationship of various query expansion models. Dinçer et al. [11] argue that unless the experimental method seeks to directly improve the reference model, it may not be reasonable to use just one baseline, especially if the baseline itself has a volatile effectiveness profile. Zhang et al. [27] apply the methods of Zhang et al. [28] to graphically evaluate the risk profiles of multiple TREC systems, and show that this can be done in an unbiased way. Dinçer et al. [9] propose an analytical method ZRisk to accommodate comparisons in terms of the risk and reward of a system against multiple baselines. A matrix of system and topic scores is used:

$$ZRisk(s_i, \alpha) = \sum_{q \in Q^+} \frac{x_{ij} - e_{ij}}{s_{ij}} + (1+\alpha) \cdot \sum_{q \in Q^-} \frac{x_{ij} - e_{ij}}{s_{ij}}, \tag{3}$$

as a form of weighted standardization [25] in which both wins and losses are scaled, with the expected values of cells based on both systems *and* topics. To normalize all of the scores to produce a fair comparison, the mean effectiveness and ZRisk are combined to produce the final result:

$$GeoRisk(s_i, \alpha) = \sqrt{Effectiveness(s_i) \cdot \Phi(ZRisk_\alpha / c)}. \tag{4}$$

Here $\Phi$ is the cumulative distribution function of the standard normal distribution, which is used to ensure that ZRisk scores are in $[0, 1]$. That is, GeoRisk values combine information about mean, variance, and shape with respect to many baselines.

## 3 Broad Issues with Trade-off Measures

The previous section discussed several quantitative risk measures. We now discuss a number of factors that are common to all of these approaches: the-user defined $\alpha$ parameter; the $(\alpha + 1)$ scalar component; and their naming.

**The $\alpha$ Trade-Off Parameter**. The $\alpha$ parameter scales the impact of losses relative to the baseline, with a range of different values employed in experimentation. Table 1 lists the $\alpha$ parameters used in a sample of ten papers that employed risk measures as part of their experimental regime. As risk evaluation goals to date have been driven more by experimental care (and caution) than by the measured experience of a cohort of users, it is unsurprising that a spread of $\alpha$ parameters has emerged, with no single value identified as the "reference" setting. Nevertheless, the use of different parameters makes comparing mechanisms a challenge across papers. A plausible solution for $\alpha$ selection, in line with the human experience of risk and reward, is to look to behavioral economics. Tversky and Kahneman [20] argue that to "break-even" in terms of perceived monetary gains and losses, individuals must earn twice as much from a "win" as they lose in a "loss", suggesting that $\alpha = 1$ be regarded as being a useful reference point. The obvious caveat here is that financial investments are quite different to IR effectiveness. User studies would need to be run to verify how closely related the user perception is of retrieval effectiveness loss to monetary loss. Similar prospect theory experiments have been carried out that explore whether gains and losses of time are perceptually similar to gains and losses of money [1, 12].

**The "Plus One" Loss Scalar**. If, as conjectured, a loss should count twice as much as a gain, one might conclude that $\alpha = 2$ should be chosen, in accordance with the way that scalar coefficients are employed in a range of other ways in

| $\alpha$ | Citations |
|---|---|
| 1, 5, 10 | Collins-Thompson et al. [6], Dinçer et al. [11], Sousa et al. [18], Benham and Culpepper [3] |
| 5 | Collins-Thompson et al. [7], McCreadie et al. [17] |
| 2 | Gallagher et al. [13], Benham et al. [4] |
| 1, 5, 10, 20 | Dinçer et al. [9] |
| 1, 2, 3, 4 | Hashemi and Kamps [14] |

**Table 1.** Differing sets of $\alpha$ employed for risk evaluation.

IR evaluation. In fact, the definitions that have evolved employ $\alpha = 2$ to mean that losses incur a *three*-fold penalty, and that $\alpha = 1$ is the correct value to use when losses have twice the cost of a similar-magnitude gain. Similar, the use of $\alpha = 0.5$ does *not* imply that losses have half the weight of gains. Given the challenge of explaining in prose how the losses are being scaled (often in the experimental sections of research papers where space is a perennial issue), users of these measures are likely to make mistakes, as are their readers. Liu et al. [16] comment on the risks associated with such "off by one" errors.

**Naming**. If the output of (for example) URisk is positive, the sum of the rewards is greater than the sum of the $\alpha$-scaled losses. That is, numerically "high" risk scores are desirable, but in English expression, have connotations that are opposite to that. Similarly, it is equally confusing (and hence "risky" in a communications sense) to have numerically low (or negative) risk score values be an indication that a new system is yielding volatile scores and needs to be treated with caution.

**Suggested Changes**. We propose that URisk be renamed to URisk$^-$ (or U$^-$), and (compare with Equation 1) be computed as

$$URisk^- = -(1/c) \cdot \left[ \sum Wins - \hat{\alpha} \cdot \sum Losses \right] . \tag{5}$$

Additionally, we suggest that TRisk be replaced by TRisk$^-$ (or T$^-$):

$$TRisk^- = URisk^- / SE(URisk^-) , \tag{6}$$

and that ZRisk be subsumed by ZRisk$^-$ (or Z$^-$):

$$ZRisk^-(s_i, \hat{\alpha}) = -1 \cdot \left[ \sum_{q \in Q^+} \frac{x_{ij} - e_{ij}}{s_{ij}} + \hat{\alpha} \cdot \sum_{q \in Q^-} \frac{x_{ij} - e_{ij}}{s_{ij}} \right] . \tag{7}$$

Finally, GeoRisk becomes GeoRisk$^-$ (or Geo$^-$), calculated as:

$$GeoRisk^-(s_i, \hat{\alpha}) = \sqrt{Effectiveness(s_i) \cdot \Phi(ZRisk^-/c)} . \tag{8}$$

In these revised definitions the signs have been reversed, and the $(\alpha+1)$ component has been replaced by a more conventional scalar, denoted as $\hat{\alpha}$ to distinguish it from $\alpha$, with $\hat{\alpha} = 1 + \alpha$. We plan to use these revised definitions in our own future work, and encourage others to also adopt them.

## 4 Smooth Value Functions

This section explores a different question – whether, if differences in effectiveness are weighted greater for outliers, meaningful changes are detected in system rankings compared to the current linearly weighted risk-sensitive models.

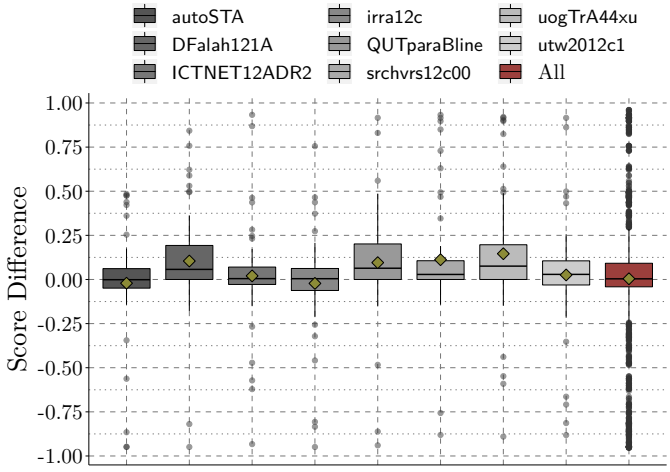


**Fig. 1.** Differences in ERR@20 for eight systems relative to the indriCASP baseline, for the TREC 2012 Web Track corpus. The "All" boxplot contains the score differences against all submitted runs to the track. Diamonds indicates the arithmetic means.

We follow Dinçer et al. [9] and past risk-sensitive evaluations, and use the run indriCASP as a baseline [6] to compute ERR@20 on the 2013 TREC Web Track corpus, where the systems evaluated against are the 48 runs submitted to the 2012 TREC Web Track. Also adopted from the methodology of Dinçer et al. [9] is the tabulation of risk comparisons against the most effective submitted run per research group, a total set of eight runs. The ERR@20 score of indriCASP is 0.195, and the median ERR@20 score of the 9 systems (top-eight runs combined with the indriCASP baseline) is 0.220 (the ERR@20 score of utw2012c1), so indriCASP is competitive among this set of champion systems.

Figure 1 shows the distribution of scores differences for each of these top-performing systems compared to indriCASP, as well as the score differences associated with all 48 systems submitted to the track. From this figure we observe that many values exceed the $1.5 \times IQR$ boundaries in the "All" boxplot, where score differences below $-0.241$ and above 0.292 are considered outliers.

**Function Definitions**. The linear piecewise function that moderates the impact of the sum of wins, minus the sum of losses, in the URisk$^-$ family of measures has the form:

$$l(x) = \begin{cases} x & x \geq 0 \\ \hat{\alpha} \cdot x & x \leq 0\,, \end{cases} \tag{9}$$

where $-1 \leq x \leq 1$ is the difference in effectiveness between baseline(s) and a run. For $\hat{\alpha} = 2$, the extrema of the domain gives the coordinates $(-1, -2)$ and $(1, 1)$. We take these two points, plus the origin, to develop an alternative continuously differentiable *smooth* weighting function. Since a user is less likely to notice small score differences [2, 19], we sought to make extreme dif-

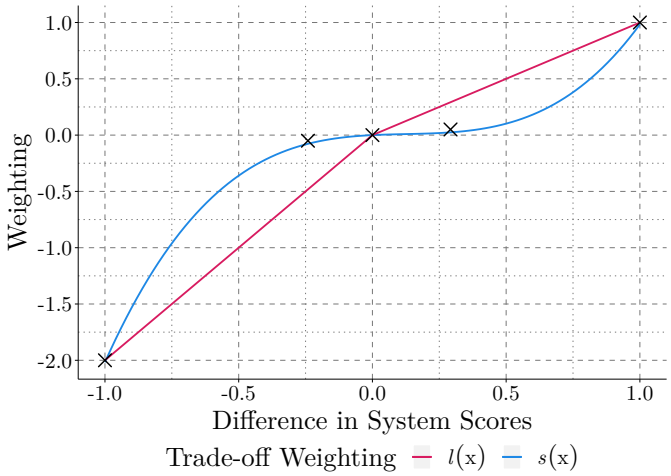


**Fig. 2.** The linear URisk$^-$ function $l(x)$ with the parameter $\hat{\alpha} = 2$, versus the proposed cubic regression variant, $s(x)$. Crosses mark the five points used to model the curve.

ferences (in both directions) count for more. Two more points, $(-0.241, -0.05)$ and $(0.292, 0.05)$, corresponding to the fences of the All boxplot in Figure 2, were taken to define a range in which the score delta was low enough to not be problematic. We then used the `R lm` function to compute a cubic regression across the five points of interest, removing the y-intercept term $d$ from the resultant function to ensure that it intersected with the origin. Figure 2 includes the result: $s(x) = 1.38426x^3 - 0.51659x^2 + 0.11578x$.

**Risk Before Standardization in ZRisk$^-$.** In order to replace $l(x)$ with $s(x)$ in ZRisk$^-$, it is necessary to first establish that ZRisk$^-$ produces the same result if the trade-off is computed before standardization. If that is the case, $s(x)$ can be computed without having to renormalize. As a reminder, ZRisk$^-$ standardizes the scores before computing the trade-off. Here we describe the scenario where we compute the trade-off and then standardize, which we call RiskZ$^-$.

Given the ZRisk$^-$ equation:

$$\begin{aligned} ZRisk^{-}(i, \hat{\alpha}) &= -1 \cdot [z_{i+} + \hat{\alpha} \cdot z_{i-}] \\ &= -1 \cdot \left[ \sum_{q \in Q^{+}} \frac{x_{ij} - e_{ij}}{s_{ij}} + \hat{\alpha} \cdot \sum_{q \in Q^{-}} \frac{x_{ij} - e_{ij}}{s_{ij}} \right] \\ &= -1 \cdot [Wins_z + Losses_z] \ , \end{aligned} \tag{10}$$

with $Wins_z$ and $Losses_z$ treated independently to show that RiskZ$^-$ is equivalent. Based on this, we can define RiskZ$^-$ as:

$$\begin{aligned} RiskZ^-\left(i,\hat{\alpha}\right) &= -1\cdot\left[\sum_{q\in Q_+} z\left(x_{ij}-e_{ij}\right)+\sum_{q\in Q_-} z\left(\hat{\alpha}\cdot\left(x_{ij}-e_{ij}\right)\right)\right] \\ &= -1\cdot\left[Wins_t+Losses_t\right]. \end{aligned} \tag{11}$$

Focusing on the $Wins_t$ part of RiskZ$^-$, we evaluate the standardized score from the standard normal distribution:

$$Wins_t=\sum_{q\in Q_+}\frac{\left(x_{ij}-e_{ij}\right)-E\left[x_{ij}-e_{ij}\right]}{s_{ij}}. \tag{12}$$

To evaluate the expected value $E\left[x_{ij}-e_{ij}\right]$, observe that the expectation operator $E\left[\cdot\right]$ is linear. Hence, $E\left[x_{ij}-e_{ij}\right]=E\left[x_{ij}\right]-E\left[e_{ij}\right]=0$, and therefore, $Wins_t=\sum_{q\in Q_+}\frac{x_{ij}-e_{ij}}{s_{ij}}=Wins_z$. In addition, $E\left[cX\right]=c\cdot E\left[X\right]$, meaning that a similar argument allows

$$Losses_t=\hat{\alpha}\cdot\sum_{q\in Q_-}\frac{x_{ij}-e_{ij}}{s_{ij}}=Losses_z. \tag{13}$$

That is, $RiskZ^-\left(i,\hat{\alpha}\right)=ZRisk^-\left(i,\hat{\alpha}\right)$.

Finally, since calculating risk before normalizing gives the same result as ZRisk$^-$, $s(x_{ij}-e_{ij})$ can be used in place of the existing linear scaling applied in ZRisk$^-$:

$$\begin{aligned} RiskZ^-{}_{s(x)}\left(i,\hat{\alpha}\right) &= -1\cdot\left[\sum_{q\in Q_+} z\left(s(x_{ij}-e_{ij})\right)+\sum_{q\in Q_-} z\left(s(x_{ij}-e_{ij})\right)\right], \\ &= -\sum_{q\in Q} z\left(s(x_{ij}-e_{ij})\right). \end{aligned} \tag{14}$$

**Smooth Cost Functions**. Wang et al. [24] proved that URisk using $l(x)$ has the property of being consistent. This is an important property for our formula, if it is to be used as a cost function in a learning-to-rank (LtR) scenario. Although we do not explicitly run any LtR experiments using the $s(x)$ cost function, we show that at least one case exists where a smooth value function can easily be used in a cost function.

The derivative of $s(x)$ is $s'(x)=4.15278x^2-1.03318x+0.11578$. Since the discriminant of $s'(x)$ is $=-0.005$, it has no real solutions, meaning $s(x)$ must be one-to-one. Moreover, since $4.15278>0$, it is clear that $s'(x)$ only returns positive values. With that in mind, combined with the knowledge that $s(x)$ is one-to-one, we have that $s(x)$ is strictly monotonically increasing; and hence no cases possible in which $d_i$ and $d_j$ could be swapped inconsistently, provided that the evaluation metric also has the property of being consistent.

Since $s(x)$ is consistent, URisk with $s(x)$ can be used as a cost function with learning-to-rank retrieval models such as LambdaMART [5].

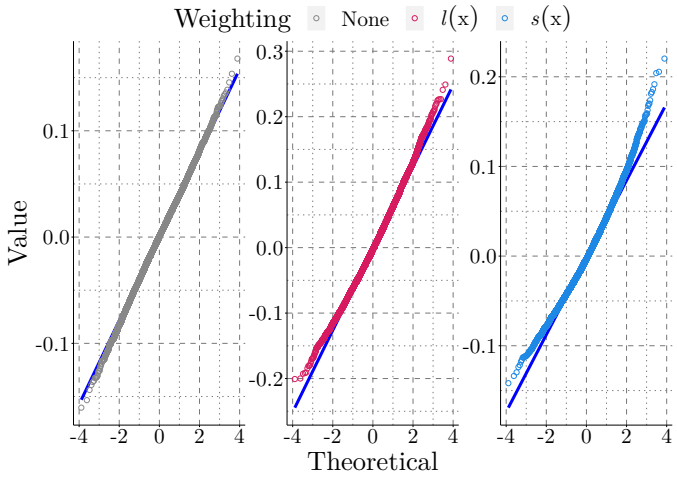


**Fig. 3.** The Q-Q plot of ICTNET12ADR3 bootstraped replicates of the mean URisk$^-$ values against the indriCASP reference system, using no risk weighting, the standard linear risk function $l(x)$ with $\hat{\alpha} = 2$, and our smooth weighting function $s(x)$.

## 5 Experiments

**Distribution Properties**. Since TRisk$^-$ is a parametric inference test, it is important to verify that the score distributions of the risk functions are amenable to statistical tests that assume normality. Hesterberg et al. [15] note that:

> *The shape of the bootstrap distribution approximates the shape of the sampling distribution, so we can use the bootstrap distribution to check the normality of the sampling distribution.*

Figure 3 shows a Q-Q plot of 10,000 bootstrapped replicates of the mean URisk$^-$ values generated with both approaches against the median scoring ERR@20 run ICTNET12ADR3, as well as a standard differences in means comparison labeled *None*, with the code to generate bootstrap replicates adapted from Urbano et al. [21]. There is evidence that the $s(x)$ score distribution has a moderate right-tail, and we flag this as a possible issue for t-test inferences since it violates the normality assumption. That could be because large differences in scores correspond to a larger mapping of the "risk" of changing to the ICTNET12ADR3 system. Despite this, the values from the $l(x)$ and no risk functions fall along their respective reference lines, providing support for the inferences made by TRisk$^-$ using these functions on the 2012 TREC Web Track dataset.

**Results**. Table 2 shows the difference in weighting functions across the top-eight systems from the 2012 Web Track.

No values fall outside the $-2.0 < t < 2.0$ "non-statistical" region when TRisk$^-$ is used with a smooth value function. As can be seen from the table, the

| System | $l(x)$ $U^-$ | $s(x)$ $U^-$ | $l(x)$ $T^-$ | $s(x)$ $T^-$ | $l(x)$ $Z^-$ | $s(x)$ $Z^-$ | $l(x)$ $Geo^-$ | $s(x)$ $Geo^-$ |
|---|---|---|---|---|---|---|---|---|
| autoSTA | 0.12 | 0.10 | 1.63 | 1.85 | 8.14 | 0.91 | −0.31 | −0.30 |
| DFalah121A | −0.05 | 0.02 | −0.85 | 0.52 | 7.02 | 0.61 | −0.41 | −0.39 |
| ICTNET12ADR2 | 0.05 | 0.03 | 0.78 | 0.75 | 6.80 | 0.17 | −0.35 | −0.33 |
| irra12c | 0.12 | 0.08 | 1.70 | 1.72 | 5.86 | 0.27 | −0.31 | −0.29 |
| QUTparaBline | −0.04 | 0.03 | −0.57 | 0.62 | 6.51 | 0.97 | −0.40 | −0.38 |
| srchvrs12c00 | −0.07 | −0.02 | −1.08 | −0.50 | 8.53 | 1.42 | −0.42 | −0.40 |
| uogTrA44xu | −0.09 | −0.03 | −1.29 | −0.63 | 5.29 | 0.70 | −0.43 | −0.41 |
| utw2012c1 | 0.06 | 0.05 | 0.79 | 1.15 | 6.33 | 0.37 | −0.35 | −0.33 |

**Table 2.** Risk values of the top-eight runs submitted to the 2012 TREC Web Track measured using ERR@20, comparing the $l(x)$ and $s(x)$ weighting functions. URisk$^-$ and TRisk$^-$ are measured for all runs against the indriCASP baseline. ZRisk$^-$ and GeoRisk$^-$ are measured using all submitted systems to the track, along with the indriCASP baseline. Shaded cells indicate the system in that row that has the least risk according to the column's measure. As the standard deviation is different for each TRisk$^-$ computation, and no t-values fall outside of the statistical region, no TRisk$^-$ cells are shaded. All risk measures use $\hat{\alpha} = 2$.

linear value function appears to have strong agreement with the smooth one when comparing risk values, suggesting that they do not generate different outcomes. And while the rankings of ZRisk$^-$ are different between the two functions, when combined as GeoRisk$^-$, the rankings are identical.

**System Risk Ordering**. Table 2, carried out on the 2012 TREC Web Track dataset, indicates that a well-behaved smooth loss-weighting function rarely changes the risk-aware system comparisons. Note, however, that only one collection and one metric were employed, and that further work is required to ascertain whether non-linear penalties provide an alternative to current piecewise-linear approaches to risk-reward analysis and LtR optimization.

To further check to see if $s(x)$ and $l(x)$ are meaningfully different loss functions, we empirically compare the similarity of the rankings generated by the URisk$^-$, ZRisk$^-$, and GeoRisk$^-$ versions of each. Ranking TRisk$^-$ t-values by different system outputs is not possible, as these values are expressed in units of $U^-$ per standard error of $U^-$, where the standard error relates to two systems only. Like Table 2, we evaluate risk values using all 48 submitted systems, where URisk$^-$ is evaluated against the indriCASP baseline [6]. For differences in orderings between $s(x)$ and $l(x)$ to have practical value, there would ideally be disagreement in the systems deemed to have the least risk. To measure the similarities of their orderings, we modulate the growth of their respective set sizes between 1 and the 49 systems (48 for URisk$^-$ excluding indriCASP) in steps of 5, and compute their similarities. As the sets produced might be non-conjoint, we cannot use popular similarity measures such as Kendall's $\tau$ or Spearman's $\rho$, and instead employ the Rank-Biased Overlap measure proposed by Webber et al. [26], since it meets our non-conjointness requirements and it can produce

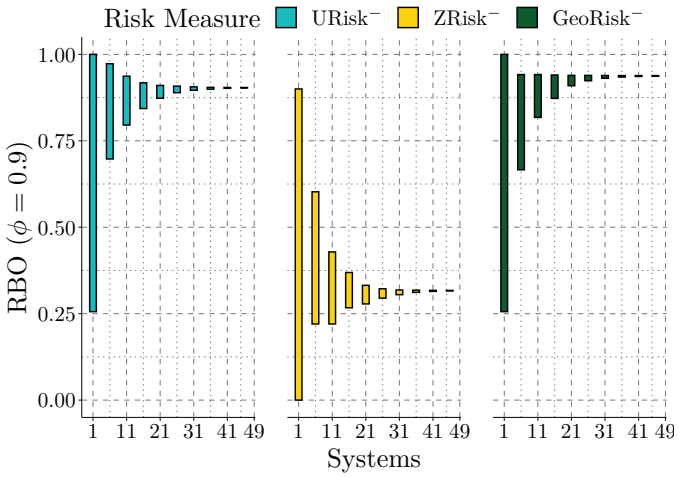


**Fig. 4.** RBO ($\phi = 0.9$), measured between $l(x)$ and $s(x)$ loss penalties, and computed over systems sets of increasing size, with the systems ordered by increasing risk scores. indriCASP is the baseline used for URisk$^-$ scores, and all systems (including indriCASP) are the baseline for ZRisk$^-$ and GeoRisk$^-$. All risk measures use $\hat{\alpha} = 2$.

top-heavy similarity scores. We fix the persistence parameter $\phi = 0.9$ for all experimentation; an RBO value of 1.0 indicates complete agreement, and an RBO value of 0.0 indicates that the two lists are disjoint.

Figure 4 shows the resultant RBO ($\phi = 0.9$) values. The lower boundary of each bar corresponds to the minimum RBO value for that set size, and the top-value is the maximum possible value, and is the lower value plus the RBO residual (the extent of the possible uncertainty as a result of the rankings being finite and not fully specified). As RBO ($\phi = 0.9$) gives an expected viewing depth of 10, we expect to see a degree of convergence of the boundaries of the RBO range as the set size increases past this point.

When all systems are considered in the RBO ($\phi = 0.9$) calculation, ZRisk$^-$ is the only measure that appears to be ranking systems very differently with the smooth value function, with a score of 0.316. But when combined with information about the mean effectiveness of the systems using GeoRisk$^-$, RBO ($\phi = 0.9$) gives a very strong similarity of 0.937 with a negligible residual. URisk$^-$ has a marginally smaller similarity score of 0.903.

## 6 Conclusion

We have explored a non-linear weighting function for IR risk evaluation measures. That function weights large differences greater than small differences in scores, on the assumption that a user is more likely to observe such changes if they occur in search results. Additionally, we proposed changes to the naming and formulae used in common risk measures, to more naturally align linguistic

conventions surrounding the terminology with mathematical interpretations of the results. In preliminary experiments with the ERR metric and TREC 2012 Web Track data, and several popular risk measures, we found no evidence to indicate that using a smooth risk function might lead to different evaluation outcomes when undertaking a risk-sensitive experimental comparison. Further work is clearly warranted, to gain a better understanding of the connection between human expectations and perception and changes in search quality, before the true value of risk-reward experimental analysis can be fully realized.

**Acknowledgments**. The first author was supported by an RMIT Vice Chancellor's PhD Scholarship. This work was also partially supported by the Australian Research Councils Discovery Projects Scheme (DP190101113).

## Bibliography

[1] Abdellaoui, M., Kemel, E.: Eliciting prospect theory when consequences are measured in time units: "Time is not money". Manag. Sci. 60(7), 1844–1859 (2013)

[2] Allan, J., Carterette, B., Lewis, J.: When will information retrieval be good enough? In: Proc. SIGIR. pp. 433–440 (2005)

[3] Benham, R., Culpepper, J.S.: Risk-reward trade-offs in rank fusion. In: Proc. ADCS. pp. 1:1–1:8 (2017)

[4] Benham, R., Culpepper, J.S., Gallagher, L., Lu, X., Mackenzie, J.: Towards efficient and effective query variant generation. In: Proc. DESIRES (2018)

[5] Burges, C.: From RankNet to LambdaRank to LambdaMART: An overview. Learning 11(23-581), 81 (2010)

[6] Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E.M.: TREC 2013 web track overview. In: Proc. TREC (2014)

[7] Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E.M.: TREC 2014 web track overview. In: Proc. TREC (2015)

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

[9] Dinçer, B.T., Macdonald, C., Ounis, I.: Risk-sensitive evaluation and learning to rank using multiple baselines. In: Proc. SIGIR. pp. 483–492 (2016)

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

[11] Dinçer, B.T., Ounis, I., Macdonald, C.: Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In: Proc. ECIR. pp. 26–38 (2014)

[12] Festjens, A., Bruyneel, S., Diecidue, E., Dewitte, S.: Time-based versus money-based decision making under risk: An experimental investigation. J. Econ. Psychol. 50, 52–72 (2015)

[13] Gallagher, L., Mackenzie, J., Culpepper, J.S.: Revisiting spam filtering in web search. In: Proc. ADCS. p. 5 (2018)

[14] Hashemi, S.H., Kamps, J.: University of Amsterdam at TREC 2014: Contextual suggestion and web tracks. In: Proc. TREC (2014)

[15] Hesterberg, T., Moore, D.S., Monaghan, S., Clipson, A., Epstein, R.: Bootstrap methods and permutation tests, vol. 5. WH Freeman & Co., New York, NY (2005)

[16] Liu, C., Yan, X., Han, J.: Mining control flow abnormality for logic error isolation. In: Proc. SDM. pp. 106–117 (2006)

[17] McCreadie, R., Deveaud, R., Albakour, M., Mackie, S., Macdonald, C., Ounis, I., Thonet, T., Dinçer, B.T.: University of Glasgow at TREC 2014: Experiments with Terrier in contextual suggestion, temporal summarisation and web tracks. In: Proc. TREC (2014)

[18] Sousa, D.X.D., Canuto, S.D., Rosa, T.C., Martins, W.S., Gonçalves, M.A.: Incorporating risk-sensitiveness into feature selection for learning to rank. In: Proc. CIKM. pp. 257–266 (2016)

[19] Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. SIGIR. pp. 11–18 (2006)

[20] Tversky, A., Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty. J. Risk Uncertain. 5(4), 297–323 (1992)

[21] Urbano, J., Lima, H., Hanjalic, A.: Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors. In: Proc. SIGIR. pp. 505–514 (2019)

[22] Voorhees, E.M.: Overview of TREC 2003. In: Proc. TREC. pp. 1–13 (2003)

[23] Wang, J., Zhuhan, J.: Portfolio theory of information retrieval. In: Proc. SIGIR. pp. 115–122 (2009)

[24] Wang, L., Bennett, P.N., Collins-Thompson, K.: Robust ranking models via risk-sensitive optimization. In: Proc. SIGIR. pp. 761–770 (2012)

[25] Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proc. SIGIR. pp. 51–58 (2008)

[26] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Sys. 28(4), 20:1–20:38 (2010)

[27] Zhang, P., Hao, L., Song, D., Wang, J., Hou, Y., Hu, B.: Generalized bias-variance evaluation of TREC participated systems. In: Proc. CIKM. pp. 1911–1914 (2014)

[28] Zhang, P., Song, D., Wang, J., Hou, Y.: Bias-variance decomposition of IR evaluation. In: Proc. SIGIR. pp. 1021–1024 (2013)