Bayesian Inferential Risk Evaluation On Multiple IR Systems

Rodger Benham
RMIT University
Melbourne, Australia

Ben Carterette
Spotify
New York, USA

J. Shane Culpepper
RMIT University
Melbourne, Australia

Alistair Moffat
The University of
Melbourne, Australia

ABSTRACT

Information retrieval (IR) ranking models in production systems continually evolve in response to user feedback, insights from research, and new developments. Rather than investing all engineering resources to produce a single challenger to the existing system, a commercial provider might choose to explore multiple new ranking models simultaneously. However, even small changes to a complex model can have unintended consequences. In particular, the per-topic effectiveness profile is likely to change, and even when an overall improvement is achieved, gains are rarely observed for *every* query, introducing the risk that some users or queries may be negatively impacted by the new model if deployed into production.

Risk adjustments that re-weight losses relative to gains and mitigate such behavior are available when making one-to-one system comparisons, but not for one-to-many or many-to-one comparisons. Moreover, no IR evaluation methodology integrates priors from previous or alternative rankers in a homogeneous inferential framework. In this work, we propose a Bayesian approach where multiple challengers are compared to a single champion. We also show that risk can be incorporated, and demonstrate the benefits of doing so. Finally, the alternative scenario that is commonly encountered in academic research is also considered, when a single challenger is compared against several previous champions.

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**; • **Mathematics of computing → Exploratory data analysis**; **Bayesian computation**.

KEYWORDS

Bayesian inference; risk-biased evaluation; multiple comparisons; effectiveness metric; credible intervals

ACM Reference Format:
Rodger Benham, Ben Carterette, J. Shane Culpepper, and Alistair Moffat. 2020. Bayesian Inferential Risk Evaluation On Multiple IR Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401033

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401033

1 INTRODUCTION

Determining how to properly compare information retrieval systems has been a central problem in the field of IR since its inception more than fifty years ago. Cleverdon et al. [10] established the foundations in which repeatable evaluation could be conducted, and despite the advent of online evaluation models for use in commercial settings, batch evaluation remains an important tool. For example, online evaluation cannot be used to evaluate alternative rankers, as click-logs may be biased to the original ranker [24]. A range of offline evaluation metrics have been proposed [25, 28, 33], including risk-sensitive overlays [38], used to avoid situations where a subset of topics gives rise to severe performance degradation, even though effectiveness is improved on average. One simple technique is to regard score degradations as having a greater weight than do score improvements of the same magnitude.

At the same time, statistical tests are widely used to rule out topic sampling error as a factor when measuring improvements [15, 30, 34, 36], usually with the goal of finding out whether a candidate ranker, or *challenger*, is demonstrably more effective than the current system – the *champion*. Sakai [31] analyzed several mistakes commonly made in IR papers, including under- and over- powered significance tests. Another issue is failing to "correct" when making multiple comparisons [19], which may lead to false discoveries when comparing many challengers, or many champions.

While applying traditional significance testing when comparing average effectiveness has proven to be beneficial in IR, it may be important to perform inference on other aspects of performance such as risk-sensitive overlays as well. Dinçer et al. [15] explore risk-sensitive inferential testing when a single champion is used as a baseline, but do not explore the multiple experimental inference case. Benham et al. [7] compare current approaches, concluding that risk-adjusted score differences can create asymmetric score distributions; that the confidence intervals from different Bootstrap approaches disagree when risk thresholds are high; and advocating the use of a biased-corrected approach. Dinçer et al. [14] note that single risk baseline evaluation biases towards the champion, and propose GeoRisk [13] as a many-to-many risk measure. GeoRisk is not an inferential measure, however, and can be difficult to interpret. In addition, publicly available collections often have *artifact* systems available, but these are rarely leveraged fully when performing system comparisons.

Bayesian inference has been shown to provide additional insights in certain scenarios as well [9, 32]. Here we explore the advantages of a Bayesian inferential approach, using past systems (and risk) to improve the reliability of multi-system comparisons. One benefit is that random effects like systems and topics shrink towards their respective mean, which makes their inferential analysis more conservative and avoids the need for multiple comparison correction [21]. Additionally, since Bayesian inference allows selection of which family the data belongs to when specifying a prior

distribution, we are able to extend the frequentist approach of Benham et al. [7] to a Bayesian one capable of accounting for skewed risk-adjusted score differences.

We present an empirical evaluation using a pool of challenger systems and a single champion, and discuss the implications for the alternative scenario of evaluating a single challenger relative to a pool of champions. More specifically, we consider these three research questions:

- **RQ1:** *How does using previous system artifacts affect Bayesian inferential results for IR test collections?*
- **RQ2:** *How does the Bayesian prior affect inferential results using risk adjusted scores, for one-to-one comparisons and one-to-many?*
- **RQ3:** *How do Bayesian and frequentist credible and confidence intervals differ when performing risk-adjusted evaluation?*

2 BACKGROUND

Risk Measures. Risk-adjusted evaluation overlays quantify the extent to which topic-specific score changes, positive and negative, might affect a comparison between a champion and a challenger system. Of interest are approaches that parameterize the weighting applied to effectiveness losses and compute a summary value. For example, URisk [38] sums differences in score while applying a weighting of r to the losses[1]:

$$URisk_r = (1/n) \cdot \left[\sum Wins - r \cdot \sum Losses \right], \tag{1}$$

where n is the number of score pairs (topics), and wins and losses are relative to the champion's effectiveness scores. Values greater than zero indicate that the challenger is more rewarding than the champion, even after scaling the losses by r. URisk has been used in evaluation [11] and as a loss function in learning-to-rank [35].

To be interpreted usefully, URisk values must be compared with risk scores for other systems, leading Dinçer et al. [15] to an inferential version of URisk, denoted TRisk. It uses values from the Student-t distribution to calculate if, after the r-weighting of losses is taken into account, the challenger is statistically better (or worse) than the champion. However, Benham et al. [7] found that the distribution of risk-adjusted scores is asymmetrical, eroding the t-test assumption of normality. Instead, Benham et al. recommend approaches that form confidence intervals on risk-adjusted scores, such as the bias-corrected accelerated Bootstrap (BCa).

Dinçer et al. [14] questioned the validity of single-champion single-challenger risk evaluation in "academic" settings, as risk-sensitivity is a characteristic of the relative system comparisons, and not an absolute property of the challenger. Since runs that are similar to the original ranker are inherently the safest choice, that bias can be avoided by forming a synthetic run composed of the average per-topic scores over a group of champions. That line of inquiry is extended by Dinçer et al. [13], with risk-sensitivity information collected for multiple runs, and used to compute ZRisk, which combines the shape and variance of scores across the set, and to compute GeoRisk, which further combines the ZRisk score with the effectiveness score. Both approaches are descriptive like URisk, and hence must be interpreted after being computed.

Benham et al. [7] note that all of these previous approaches lead to high numeric values when risk is "low". To avoid ambiguity, we adopt their suggestion and multiply each score by -1 so that high values correspond to high risk, with the change indicated by a "–" appended to the method's name, to get URisk$^-$ and so on.

Bayesian MCMC. Kruschke [26] explains the two key ideas in a Bayesian data analysis: "*(1) Bayesian inference is reallocation of credibility across possibilities*"; and "*(2) The possibilities, over which we allocate credibility, are parameter values in meaningful mathematical models*". These are captured in Bayes' rule:

$$p(\theta \mid d) = \frac{p(d \mid \theta) \cdot p(\theta)}{p(d)}, \tag{2}$$

where θ is the set parameters (or hypothesis); d is the data (or evidence); $p(\theta \mid d)$ is the posterior distribution; $p(d \mid \theta)$ is the likelihood; $p(\theta)$ is the prior; and $p(d)$ is, as Lambert [27] explains: "*the probability distribution for a future data sample given our choice of model*". The posterior distribution is composed of combinations of parameters that form credible generations of the underlying data.

Carterette [9] explored an approach that models system relevance directly from the judgments themselves instead of effectiveness metrics. That approach was extended by Sakai [32] to include an effect size computation using Glass' Δ.

The family of methods that support modern Bayesian inference are known as *Markov Chain Monte-Carlo* (MCMC) simulations. Lambert [27] observes that it is usually impossible to independently sample from the global posterior distribution, but that it can be done locally. Figure 1 gives a high-level overview of the generalized case when conducting a Bayesian MCMC experiment. The inputs are a data-set d; prior beliefs θ about the data; and a set of instructions for the sampling algorithm. How θ is organized depends on the trends in d, whether there are hierarchical attributes to model; and how informative or weakly-informative (the data has a heavy-emphasis on the outcome) it should be. Similarly, the number of chains, iterations, warmup iterations, and the seed used all play a role in determining the outcome of the Bayesian posterior produced. In the *warm-up* period an initial set of iterations are trimmed from the final result, to discard iterations that had not yet converged upon the posterior distribution. Lambert [27] recommend 4–8 chains for simple models, and "few tens of chains" for complex models, to ensure the MCMC sampler does not settle in local minima. Many sampling algorithms exist, and the state-of-the-art approach is the No-U-Turn Sampler (NUTS) [23], implemented in the Stan language.

Many heuristics exist to ensure that the chains have mixed, including $\hat{R}$ [20], where values closest to 1.0 are optimal. A full suite of diagnostic tools is provided in `shinystan` in R [29]. Because sampling methods at different points of the simulation can fail to meet the condition to sample more of the posterior, Lambert [27] suggest that an *effective sample size* is used to describe how many dependent samples are required to interpret statistics describing the posterior, mapped to a value of how many independent samples there are. For estimates of the 95% high-density interval limits, Kruschke [26] recommends an effective sample size of at least 10,000. If there are no diagnostic issues, the posterior may then be used.

[1] To simplify discussion, the original $\alpha + 1$ losses weighting [38] is folded into r. For example, $r = 2$ indicates a two-fold scaling of losses relative to gains [7].

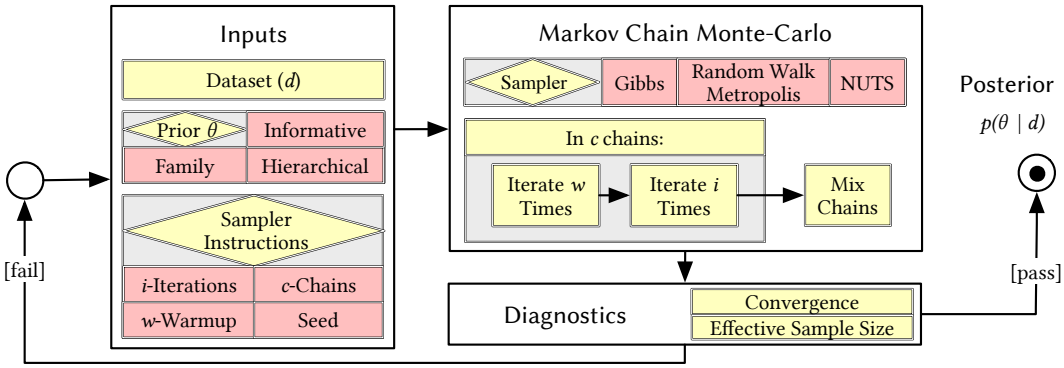

Figure 1: Bayesian Markov-Chain Monte Carlo. Researcher decisions are indicated by diamonds, with the choices as red rectangles. Note that this is a high-level diagram and the diagnostics listed are not exhaustive.

Factor Analysis. A range of recent results in regard to variance contribute to understanding how to reduce error when modeling scores. Ferro and Silvello [16] use a *General Linear Mixed Model* (GLMM) to approximate system performance as an amalgamation of system effect, topic effect, and system-topic interactions, establishing the relative impact on retrieval scores of stop lists, stemmers, and retrieval models. GLMMs and two-way ANOVA were also used to explore shard and topic-shard interaction effects in distributed retrieval, greatly reducing the regression error [17]. Carterette [8] compares a Bayesian linear model against the t-test, noting that using a t-test means implicitly accepting a model too.

In our work here we employ aspects of all of these to model Bayesian inferences with topic and system effects, but excluding their interaction (system-topic effects, in other words) and shard effects in order to make simulation speeds tractable. Those component effects might be added back in the future to improve the reliability of score point estimates and credible intervals.

3 BAYESIAN HIERARCHICAL MODELING

We employ a hierarchical model for system and topic effects:

$$\begin{aligned} y_{ts} &= \beta_0 + T_{0t} + S_{0s} + e_{ts}, \\ T_{0t} &\sim \mathcal{N}(0, \tau_{00}^2), \\ S_{0s} &\sim \mathcal{N}(0, \chi_{00}^2), \\ e_{ts} &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3}$$

where y_{ts} is the effectiveness score observed for system s and topic t; T_{0t} is the effect size of topic t; S_{0s} is the effect size of system s; and e_{ts} is an error term. Since other IR tests assume normality, we use a simple Gaussian prior distribution in this example, selecting the mathematical objects that describe the distributions, and leaving their hyper-parameters τ_{00}^2, χ_{00}^2, and σ^2 to be automatically determined during the simulation, based on the input data.

In order to find the most appropriate challenger to replace the champion, Bayesian sampling is used in an MCMC simulation. One advantage of sampling from the posterior is the generation of *credible intervals* for almost any summary statistic, as well as predictive intervals on data observations. Figure 2 shows the relationship between these various components.

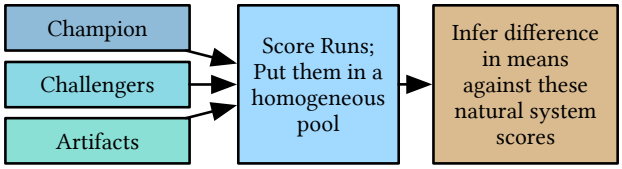

Figure 2: The inferential evaluation methodology. Information from previous systems (artifacts) is combined with runs, to determine whether to switch from the current system to any one of a set of challenger systems.

System	301	306	311	316	321	Mean
Champion	0.05	0.21	0.48	0.62	0.29	0.33
Challenger 1	0.06	0.24	0.42	0.62	0.34	0.34
Challenger 2	0.06	0.24	0.43	0.62	0.34	0.34
Challenger 3	0.04	0.19	0.46	0.62	0.30	0.32
Challenger 4	0.19	0.09	0.32	0.65	0.34	0.32

Table 1: System-topic scores used to demonstrate the utility of the Bayesian hierarchical inference model.

Table 1 lists system-topic scores for an example over five systems and five topics, with a champion system being compared against four challengers. As well, a sample of 79 artifact systems (not shown) over the same test data is used to generate a posterior distribution. Figure 3 shows the result of modeling the system-topic scores using sampling. Each observation has an associated predictive interval generated out of the simulation, and each model parameter has a credible interval. With only five different scores associated with each system, it makes sense that no clear winner (nor loser) emerges. The context provided by the artifact systems makes it apparent that all of the observed scores are in line with what is expected.

The Bayesian model is based on the hierarchical model of Gelman et al. [22]. The key idea is to use a "common population distribution" as a *hyper-prior* with learned *hyper-parameters* as we did in Equation 3 – for example, $\mathcal{N}$ is the hyper-prior of T_{0t}, which is

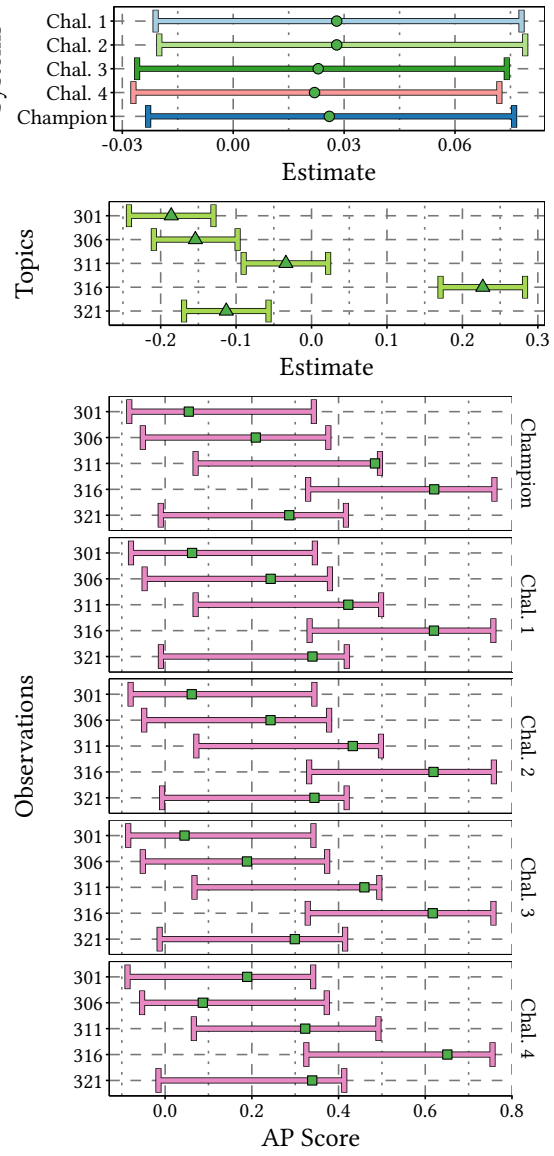

Figure 3: Granularity of information made available when performing Bayesian inference. At top, credible intervals for the system effect intercepts are plotted with a green circle indicating the random effect point; and below it topic effect intervals are shown in green, with point estimates represented by triangles. Forecasts for 95% intervals based on the score observations are shown in pink in the remaining part of the diagram, with squares corresponding to actual scores (Table 1).

composed of the hyper-parameters 0 and τ_{00}^2. When the generalized linear model is created, with results for different instantiations of a single system grouped together and ordered topic-wise, the resulting intervals have subtle variations as performance changes, and at the same time converge towards the average effect (a concept referred to as *partial-pooling*). If full-pooling is used, a single average represents that random effect. When no pooling is used in a non-hierarchical model, prediction over-fitting becomes likely, and hence the model will not generalize to new data in the form of unseen systems. Intuitively, the convergence makes Bayesian inference more conservative when trying to find statistically significant random effects (for example, the challenger systems in Table 1), unless a highly credible effect exists.

As an aside, these results point to a potentially interesting effect in regard to the evaluation bias of runs that contributed to relevance judgments for a collection. Since runs that did, and did not, contribute to the judgment set are used as artifact systems, bias can be somewhat alleviated through partial pooling.

System	Description	Robust04	TREC17	TREC18
Champion	BM25	0.274	0.210	0.236
Chal. 1	DPH + Bo1	0.323	0.289	0.301
Chal. 2	DPH + DRF + SD + Bo1	0.322	0.290	0.300
Chal. 3	DPH + DRF + SD	0.264	0.216	0.231
Chal. 4	Top-3 TREC runs fused	0.380	0.572	0.459

Table 2: Experimental systems and their average precision (AP) scores over the three different collections.

4 EXPERIMENTS

Section 3 illustrated the power of Bayesian hierarchical modeling via five systems and five topics. This section examines a more detailed – and more typical – experimental configuration, using three different document collections, and 50 topics for each.

Runs. The champion and first three challengers are from the Terrier v5.2 search engine, taking configurations originally proposed in the *SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)* workshop [2] as representative retrieval models. Table 2 describes these runs, and their corresponding mean-over-topics effectiveness computed on three collections[2]. Bo1 query expansion was included in the second challenger, with ten feedback documents used, and 25 terms selected.

The fourth challenger was designed to be a clear winner, and was produced as a fusion run merging the best three participant runs submitted to the TREC rounds associated with the three document collections. These nine runs (three per collection) were omitted from the set of artifact systems used to compute the Bayesian prior. For Robust04, CombSUM [18] fusion was used; and for TREC17 and TREC18, Reciprocal Rank fusion (RRF) [12] was employed. Both of these are simple techniques known to generate effective outcomes. All of the five runs are interesting in a risk-sensitive retrieval context, as query expansion is known to be risky due to query drift, and rank fusion has had low risk in previous comparisons [6].

Corpora. Three collections of news documents were employed: the classic Robust04 corpus, and the more recent *New York Times* and *Washington Post* collections, TREC17 and TREC18 respectively.

- Robust04 [37]: Every fifth topic was used 301, 306, and so on, to obtain 50 topics. The collection contains ≈ 528k documents, ≈ 664k unique terms, and ≈ 253M terms in total. Of the 110 runs

[2]See *Terrier v5.2 Documentation* for details of these options and settings, https://github.com/terrier-org/terrier-core/blob/5.x/doc/index.md, accessed 22 January 2020.

submitted, $m = 79$ are available as artifacts, once the bottom 25% had been removed, and the top three removed to create Challenger 4.
- TREC17 [1]: The TREC CORE 2017 Track exercise, with ≈ 1.85M documents, ≈ 2.97M unique terms, and ≈ 1.28T terms in total. There were 75 runs submitted, with $m = 53$ after removing the bottom 25% of runs and the top three.
- TREC18: The TREC CORE 2018 Track exercise, ≈ 595k documents, ≈ 1.47M unique terms, and ≈ 481B terms in total. Of 72 runs submitted, $m = 51$ were used as artifacts.

Bayesian MCMC. All simulations were executed using 12 Markov chains for 12,000 iterations, after 6,000 iterations of warm-up. Posteriors with 72,000 draws were produced when the chains were mixed. These choices were informed by inspecting the effective sample size of the posteriors post-hoc, ensuring that they were all above 10,000 to support inference with a 95% credible interval. To aid in reproducibility, we report the random seed value of 12,345.

For transparency, all MCMC involving only effectiveness measures was computed with `rstanarm`, as extra families are not needed. If an experiment involved computing system scores with risk-adjusted scores applied, we used `brms` for either Gaussian or skew-normal prior experimentation. Both packages interface with the Stan probabilistic programming language. Both Stan front-ends use the `lmer` syntax when specifying a model, which is `score ~ (1 | system) + (1 | topic)` [5]. Only Gaussian hyper-priors have been applied in hierarchical modeling previously, as selecting other distributions could reduce the ability to converge to the grand mean of the random effects, resulting in over-fitted models[3].

Metric. Runs and risk overlays were scored using average precision (AP), the official metric for all three of the collections. The first three experiments use AP scores without risk-adjustment. Our choice of risk-parameter is $r = 5$ for tabulated analysis, a moderate–high value from the set of conventional IR values: 1, 2, 5, 10. When showing the effect of the risk parameter in diagrams, we adhere to this norm [13, 15, 38].

Inference over Multiple Systems. Having provided details of the experimental resources and settings, we now describe the sequence of results obtained using them, comparing the systems in Table 2.

Figure 4 shows the benefit of employing all of the elements shown in Figure 2 (with their exclusion policies) for the scenario in which the champion is compared against the four challengers. For all three corpora the champion system can be completely differentiated from *Challenger 4* (which was the expected outcome), but with the Bayesian approach being more conservative than its frequentist counterpart. The lower plots in Figure 4 show the 95% credible intervals for topic scores, and demonstrate the ability of the Bayesian approach to differentiate between "easy" and "hard" topics, opening avenues for possible future work.

The Effect Of Artifact Systems. Recall (Figure 2) that a set of artifact systems is used as background information. To explore the way in which the number of artifacts affects the discriminative power of the inferences generated, the number of artifact systems included was varied, after first ordering them from highest to lowest score. This is motivated by the concerns raised by Armstrong et al. [3], where one of the main messages is that it is important to evaluate against strong baselines. The number of systems in the pool in all cases is $m + 5$, with m varying as the set of artifacts is extended. Note that in order for the total of $m + 5$ systems to be fairly compared, an accurate prior distribution is critical, and hence changing the number of challengers would also have an effect the outcome of the inferences. Nevertheless, when the number of artifacts is substantially greater than the number of competing systems, the outcomes can be expected to be relatively stable, and here the champion and four challengers are compared without varying the number of challengers.

Figure 5 shows the outcome for the Robust04 and TREC17 collections (TREC18 omitted, but with results very similar to those for TREC17), comparing the champion to *Challenger 4*. When $m = 1$, the champion system and *Challenger 4* are barely separable via a 95% credible interval for Robust04. When $m \in 5, 10, 20$, the systems are more distinguishable from each other; but as m approaches 40 the gap between the credible intervals starts to decrease again. This suggests that we cannot expect statistical significance to monotonically increase with respect to m as weaker systems are introduced to the pool of artifacts, a somewhat surprising result. The TREC17 collection yields a somewhat different pattern – there is a greater gap between the credible intervals at $m = 40$ than at $m = 5$. We conjecture that this effect is evidence of the high quality of runs submitted to the Robust04 track. Since effective artifact runs are being referenced as our universe of valid systems and our runs are also effective, more variability in effectiveness scores must be shown to differentiate runs from others.

One concern with this experiment is that we are knowingly including less effective systems, with the systems added in most-effective to least-effective order. This might suggest that the ability to discriminate good and bad runs increases because the prior is being conditioned on lower quality runs. However, in many cases in which the properties of runs are being exhaustively explored, the bottom 25% of runs are discarded. Since we wish to compare this approach to that of frequentist approaches (which are free to include in their methodology system runs that may not be the most competitive in real-life), we use all available run information, along with the most sensible exclusions applied as described above. Simulation studies such as those of Urbano et al. [36] may provide additional insights to establish the power of these approaches, an area that we leave for future work. (Note that the goal of this paper was to use real systems for evaluation where possible, and to develop the complete Bayesian framework.)

We are now in a position to answer **RQ1**: the number of system artifacts used in a Bayesian inferential evaluation scenario plays an important role in determining the quality of the generated models, and the more runs available, the more consistent the performance.

Academic Evaluation. In a typical academic evaluation scenario, a researcher will propose a new retrieval model as a challenger, and needs to show improvement relative to a set of champions, rather than comparing against a single system. In the Bayesian framework we have described, since all champion and challenger runs are input into a single inferential framework, the run labels

[3]Explained in more detail at https://github.com/paul-buerkner/brms/issues/231, accessed 22 January 2020

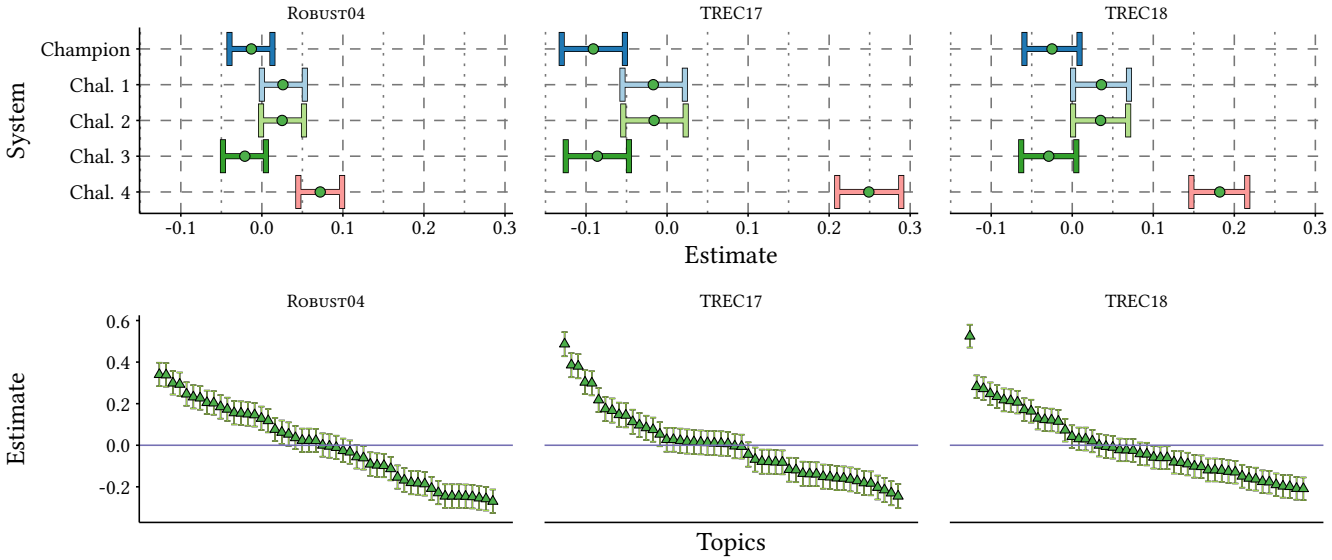

Figure 4: The full version of Figure 3, comparing five systems (Table 2) over 50 topics and three test collections. Top: the champion and *Challenger 4* can be differentiated in each case (positive score differences are desirable). Bottom: the sets of fifty credible topic score intervals.

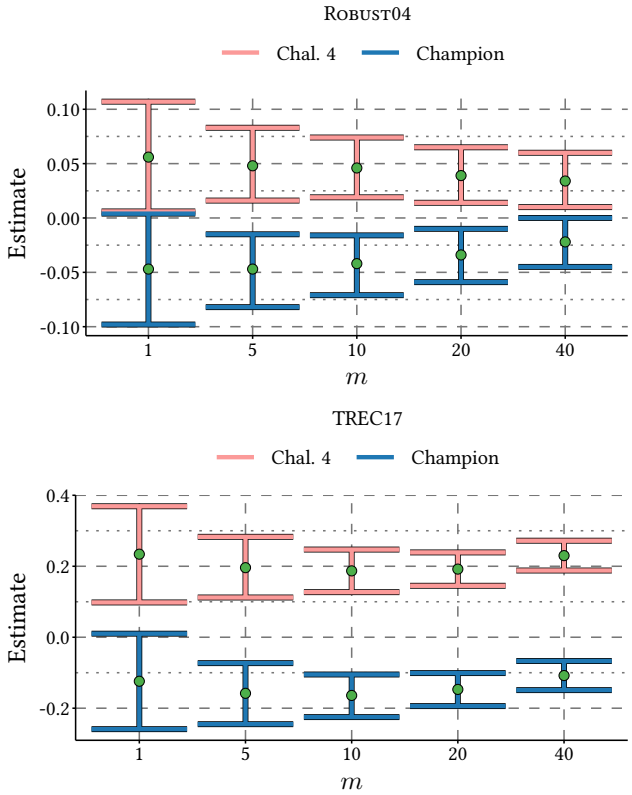

Figure 5: Discrimination between the champion and *Challenger 4* improves as the number of artifact systems increases.

can be permuted as required, and the credible intervals checked to determine if the desired outcomes hold. (This is not the case for risk overlays considered shortly, as risk-adjusted score differences are developed using a single champion as a reference point).

The Bayesian approach is also flexible enough to support many-vs-many system comparisons too. Suppose in Table 2 that we had originally set out to show that *Challenger 4* is more effective than the other systems when regarded as a set of champions. In Figure 4, we would see for TREC17 and TREC18 that it is indeed the best system. On Robust04, however, the credible interval overlaps on both *Challenger 1* and *Challenger 2*.

A key piece of what makes Bayesian inference over multiple artifact systems attractive is its capacity to give robust inferences that are non-committally tied to the set of baselines being evaluated against. Since χ_{00} in Equation 3 corresponds to the grand variance over the $m + n + 1$ systems used in the system evaluation pool – where m can be regarded as being sufficiently large – it matters little to the outcome of the evaluation if a specific baseline system was or was not evaluated against, provided that a baseline with similar mean effectiveness *was* included. That is, a specific comparison relative to one recent result or another is no longer an experimental requirement, as long as some of the systems in the comparison pool are high quality and state-of-the-art. So concerns about baselines not being a very specific recent result are no longer as important when evaluating in this framework. Evidence of this claim is visible in all of the system comparison graphs in this paper, with the credible intervals for all system effects nearly identical.

Bayesian Risk (One-To-One). Other work has shown that, as risk overlays linearly scale losses, the distribution becomes one-sided and may affect statistical methods that assume normality [7].

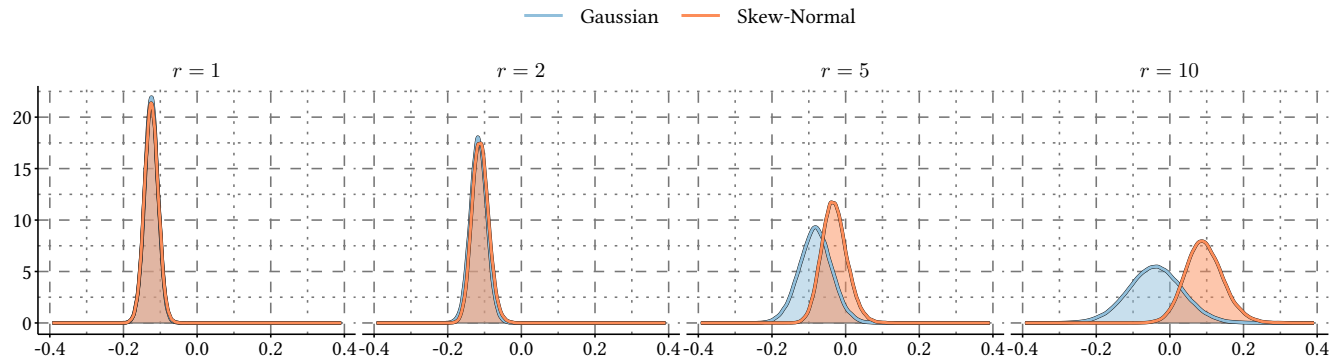

Figure 6: Density plots for URisk^{-} as r is increased in a Bayesian MCMC simulation, showing the density of the mean for each prior distribution family for $r \in \{1, 2, 5, 10\}$. A BM25 baseline run is compared against the best run `pircRB04td2` submitted to Robust04 as measured by AP, with topics sub-sampled to a traditional evaluation of 50 topics from a total of 249 available.

Returning to a single system versus single system framework, Figure 6 shows the distribution of mean scores for all of the 72,000 draws available from the posterior when risk-adjusted score differences are modeled as a Gaussian distribution, compared to a skew-normal distribution. These results confirm that the distribution becomes one-sided as r (the risk parameter) increases: the difference at $r = 1$ and $r = 2$ is negligible; but for $r = 5$ and $r = 10$ the intervals shifts to the right. In the frequentist paradigm the confidence intervals simply become more extended, and provide lower confidence predictions. Here, the density of the location parameter for the skew-normal distribution is narrower than the Gaussian alternative, a consequence of the data being more skewed after the score differences have been adjusted to the $r = 5$ and $r = 10$ levels.

The inferential effects are due to the risk overlay and not reliant on any particular run or corpus combination. We also expect that increasing the number of topics evaluated to 249 will left-shift the skew-normal distribution to agree more with the Gaussian one, due to the central limit theorem. Although the skew-normal prior has proved useful in a one-to-one comparison, as noted in Section 3, the hyper-prior must be Gaussian, and so the outcome of this part cannot be used for multiple system evaluations of risk, providing at least a partial answer to **RQ2**: the one-to-one prior can be safely set to a skew-normal distribution, where it has a tighter density plot with respect to the location parameter. However, due to current constraints with multi-level modeling, risk inferences in one-to-many evaluation contexts must use a Gaussian distribution, which may give less consistent results for high r values. Nevertheless, evaluating risk with multiple systems is still possible subject to certain caveats, as is considered shortly.

Bayesian Risk (One-Versus-Many). Extending the one-versus-many Bayesian approach to include a risk overlay is relatively simple, provided that it is remembered that absolute system scores are being considered, rather than score differences. To accomplish that, the champion's score are retained unchanged, with no score adjustments applied; while all challenger and artifact systems have risk-adjusted score differences applied to their per-topic scores. For example, if the champion's score for some topic was 0.50, and a challenger system obtained 0.45, an $r = 2$ adjustment would see that second score modified to 0.40 before inclusion in the simulation.

A natural mathematical distribution to represent this score distribution is the skew-normal [4]. Many system scores must be parameterized by a Gaussian hyper-prior, however, and this approach will only be useful when forming bounds on observations. That distribution is parameterized by a shape parameter as well as location and scale, which can be simulated using the Stan-driven Bayesian MCMC `brms` package.

Figure 7 compares the boundaries formed on system score observations for the skew-normal and Gaussian distributions when modeling risk, computed by including the sets of artifact runs (with their exclusion policies) when comparing the system effectiveness of a champion against the 4 challengers, for a range of values of r. This figure shows only Robust04 results, but system orderings by risk are consistent with those of TREC17 and TREC18. When $r = 1$, no risk is applied and the x-axis corresponds to (a reflection of) the y-axis in Figure 4. At $r = 2$, *Challenger 4* remains more risk-sensitive than *Challenger 3* when compared to the champion system. When $r = 5$, the credible intervals begin to overlap, and they remain that way for $r = 10$. When $r = 10$, the system effect size point estimate of the champion system catches up to *Challenger 1* and *Challenger 2*, while the point estimate for *Challenger 4* remains marginally superior. This sequence of data provides Bayesian support for the result of Benham and Culpepper [6], that fusion reduces risk.

In Section 3 we commented on the utility of modeling risk-scores using a skew-normal distribution rather than a Gaussian one for topic-by-topic comparisons. As *Challenger 3* was consistently the riskiest choice, exploring which topics failed against the expected values learned from the background distribution of system scores might be illuminating. Figure 8 compares the observed score from the champion system when compared against the risk adjusted score for *Challenger 3*, using $r = 5$. Scores in the positive direction indicate risk; negative scores, reward.

On the skew-normal graph, the score for the champion system gives the highest reward across all predictive forecasts most of the time, indicating that the model is fitting the data well, since its scores are not penalized. On the other hand, the Gaussian plot is

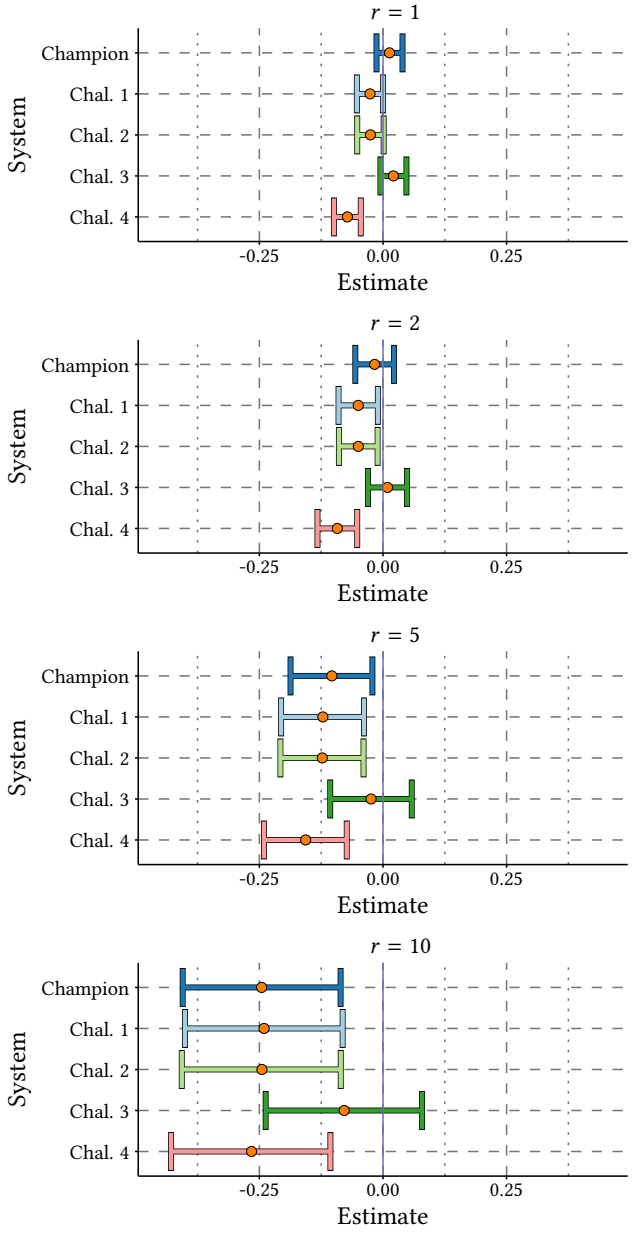

Figure 7: Inferential risk comparisons over multiple systems, with point estimates in orange. Note that the definition of risk used here means that "leftward" relationships indicate "rewarding" outcomes. *Challenger 4* is the best system to use without any risk adjustments applied, but for $r = 5$ onward it becomes indeterminate against the champion and the other challengers.

unable to resolve lower score values particularly well, conflating the original system scores with the risk-adjusted scores. Observed risk adjusted scores from the *Challenger 3* run are consistently higher across topics, due to it being riskier on average than the champion run. Of particular interest are the two riskiest topics, which have risk values so high that they sit right on the boundary of the 95%

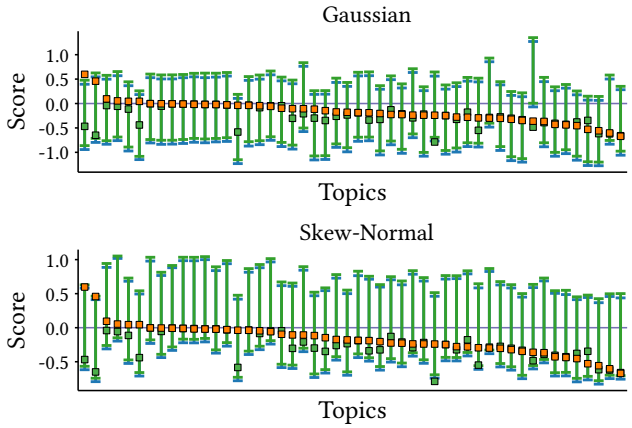

Figure 8: Per-topic risk observations for a Gaussian with a skew-normal prior compared using a 95% predictive interval for a champion system (blue) with scores represented in green, relative to *Challenger 3*, shown with a green predictive interval and an orange point marking the risk adjusted score.

skew-normal forecast of where the score should lie. While there are topics on which *Challenger 3* performs marginally better, nothing can be decided statistically as to which topic is less risky than any of the champion run's scores when $r = 5$. We now answer **RQ2**, having presented a complete example of how the Bayesian prior impacts the results of Bayesian risk data analysis.

Bayesian Risk In Context. We have discussed at length how Bayesian techniques can be used to model effectiveness and risk. We now place the Bayesian risk discussion into context with existing measures that provide a numeric output and sometimes an accompanying frequentist inference. The goal remains the same; to compare the risk of one champion system versus many challengers. For methods that compare one system against another, we explore the use of URisk$^-$, TRisk$^-$, and BCa$^-$. TRisk$^-$ scores are not corrected for multiple comparisons, consistent with the original paper, and are useful as a point of reference to show the "false" power that is present when the correction is not done. BCa$^-$ risk intervals are corrected using Bonferroni, with four hypotheses. We introduce the only approach designed to support one-versus-many comparisons as BRisk$^-$, or B- for short. Finally, ZRisk$^-$ and GeoRisk$^-$ are included as many-vs-many measures. We do not include the m artifact systems into the pool of systems for these many-to-many risk measures, as it appears that a large number of systems reduces the resolution between the scores, which become skewed due to the loss penalty. In our approach, we use a large pool of artifact systems which are risk-adjusted to improve the signal of the Bayesian approach, and to help distinguish boundaries on observed risk values.

Table 3 shows the results of this all-in experiment, over the three document collections. The inferential risk overlays that compare one system against another – with or without Bonferroni correction – are more likely than BRisk$^-$ to declare that changing to an alternative system is either statistically rewarding or statistically risky. That happens because BRisk$^-$ is constrained in the inferences

	Run		One vs. one			One vs. many	Many vs. many	
	System	AP	URisk$^-$	TRisk$^-$	BCa$^-$	BRisk$^-$	ZRisk$^-$	GeoRisk$^-$
Robust04	Champion	0.274	–	–	–	−0.103 [−0.187,−0.021]	12.42	−0.405
	Challenger 1	0.323	−0.024	−1.408	−0.024 [−0.067, 0.020]	−0.122 [−0.206,−0.038]	10.23	−0.433
	Challenger 2	0.322	−0.026	−1.581	−0.026 [−0.066, 0.016]	−0.123 [−0.207,−0.039]	9.31	−0.430
	Challenger 3	0.264	0.105	2.976	0.105 [0.042, 0.236]	−0.024 [−0.107, 0.058]	11.04	−0.394
	Challenger 4	0.380	−0.071	−2.345	−0.071 [−0.128, 0.031]	−0.156 [−0.241,−0.073]	10.65	−0.471
TREC17	Champion	0.210	–	–	–	0.017 [−0.065, 0.099]	26.03	−0.383
	Challenger 1	0.289	−0.052	−2.077	−0.052 [−0.100, 0.031]	−0.025 [−0.107, 0.057]	23.03	−0.443
	Challenger 2	0.290	−0.053	−2.114	−0.053 [−0.100, 0.034]	−0.025 [−0.107, 0.057]	22.80	−0.443
	Challenger 3	0.216	0.015	1.817	0.015 [−0.001, 0.043]	0.029 [−0.052, 0.110]	25.97	−0.388
	Challenger 4	0.572	−0.352	−11.100	−0.352 [−0.419,−0.257]	−0.263 [−0.349,−0.179]	23.56	−0.624
TREC18	Champion	0.236	–	–	–	−0.116 [−0.207,−0.024]	17.45	−0.387
	Challenger 1	0.301	−0.040	−1.882	−0.040 [−0.091, 0.014]	−0.151 [−0.244,−0.060]	16.12	−0.434
	Challenger 2	0.300	−0.042	−2.047	−0.042 [−0.092, 0.008]	−0.153 [−0.245,−0.061]	15.39	−0.432
	Challenger 3	0.231	0.065	3.468	0.065 [0.027, 0.123]	−0.058 [−0.151, 0.033]	18.93	−0.387
	Challenger 4	0.459	−0.165	−2.791	−0.165 [−0.266, 0.071]	−0.261 [−0.354,−0.169]	19.75	−0.548

Table 3: Comparison of four challenger systems against a champion system using existing risk measures, and compared against the proposed Bayesian risk measure BRisk$^-$, using $r = 5$ throughout. The BCa$^-$ interval is Bonferroni corrected for multiple comparisons, while TRisk$^-$ is not, to be consistent with the original publication, and as a point of reference to give an intuition behind the loss of power when correction in applied. Positive values in blue indicate that a system has a statistically significantly higher risk compared to the champion; negative values in blue indicate significant reward when compared to the champion.

it produces, where simulated effectiveness scores must be sane with respect to the pool of systems provided to the model, thereby answering **RQ3**. In comparing the outputs of the many-to-many risk measures, GeoRisk$^-$ always preferences the highest-scoring run. Using only ZRisk$^-$ as a guide, *Challenger 2* is the system with the lowest risk over every collection. That suggestion may be biased to the runs supplied to ZRisk$^-$ however, as *Challenger 1* and *Challenger 3* both share components with *Challenger 2* whereas the BM25 champion and the fusion run are penalized like the single baseline case discussed in Dinçer et al. [14].

The Bayesian approach is capable of leveraging additional systems as artifacts to produce better informed decisions than before – where confidence in an evaluation stems from known system effectiveness profiles. Also, exploratory and failure-based data analysis over topics and systems is fully supported by the new framework, as long as there is a sufficient signal.

It should be noted that the above framework does not invalidate the results of other testing methodologies, as no complete gold standard list of *certain* inferences exists to refer back to, in determining the correctness of statistical tests for different sample statistics. Running simulation studies to compare the false positive rate of this approach to others in the standard IR way [34, 36] would be very expensive indeed. As the definition of what a probability value should represent is subjective, an objective researcher will report both Bayesian and frequentist inferences, and base their decisions on multiple forms of evidence.

5 CONCLUSIONS

We have applied an adaptation of Bayesian hierarchical modeling using MCMC to the task of comparing the effectiveness of multiple retrieval systems. Using that methodology, we explored inferential evaluation of topic and system data comparing a champion system to a set of challengers, both without and with the overlay of risk.

In a study using run configurations from the RIGOR workshop, we found that including many diverse runs as artifacts enables the Bayesian inferential approach to be both discriminatory and generalizable. Given the nature of inferring information from past experience, the outcome of these experiments would naturally change as more evidence about the universe of known systems grows. We also explored the use of a skew-normal prior instead of the Gaussian when evaluating risk adjusted scores, and found that it was useful for one-to-one evaluation scenarios, or when forecasting risk score observations in an exploratory topic-wise data analysis.

Finally, we found that for the systems used in this study, Bayesian and frequentist confidence and credible intervals differ, and that Bayesian Risk (BRisk$^-$) is conservative in its inferential statements, based on pools of scores that are produced by a set of legitimate artifact IR systems.

Software. Code that reproduces the experiments reported in this work is available at https://github.com/rmit-ir/bayesian-risk.

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

REFERENCES

[1] J. Allan, D. Harman, E. Kanoulas, D. Li, C. Van Gysel, and E. M. Voorhees. TREC 2017 common core track overview. In *Proc. TREC*, pages 1–14, 2017.

[2] J. Arguello, F. Diaz, J. Lin, and A. Trotman. SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). In *Proc. SIGIR*, pages 1147–1148, 2015.

[3] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proc. CIKM*, pages 601–610, 2009.

[4] A. Azzalini. A class of distributions which includes the normal ones. *Scand. J. Stat.*, 12(2):171–178, 1985.

[5] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.*, 68(3):255–278, 2013.

[6] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. ADCS*, pages 1:1–1:8, 2017.

[7] R. Benham, B. Carterette, A. Moffat, and J. S. Culpepper. Taking risks with confidence. In *Proc. ADCS*, pages 1–4, 2019.

[8] B. Carterette. Model-based inference about IR systems. In *Proc. ICTIR*, pages 101–112, 2011.

[9] B. Carterette. Bayesian inference for information retrieval evaluation. In *Proc. ICTIR*, pages 31–40, 2015.

[10] C. W. Cleverdon, J. Mills, and M. Keen. Factors determining the performance of indexing systems. *Cranfield: College of Aeronautics*, 1(28), 1966.

[11] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2013 web track overview. In *Proc. TREC*, 2014.

[12] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proc. SIGIR*, pages 758–759, 2009.

[13] B. T. Dinçer, C. Macdonald, and I. Ounis. Risk-sensitive evaluation and learning to rank using multiple baselines. In *Proc. SIGIR*, pages 483–492, 2016.

[14] B. Dinçer, I. Ounis, and C. Macdonald. Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In *Proc. ECIR*, pages 26–38, 2014.

[15] 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.

[16] N. Ferro and G. Silvello. A general linear mixed models approach to study system component effects. In *Proc. SIGIR*, pages 25–34, 2016.

[17] N. Ferro, Y. Kim, and M. Sanderson. Using collection shards to study retrieval performance effect sizes. *ACM Trans. Inf. Sys.*, 37(3):30, 2019.

[18] E. A. Fox and J. A. Shaw. Combination of multiple searches. *Proc. TREC*, pages 243–252, 1994.

[19] N. Fuhr. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41, 2018.

[20] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Sci*, 7(4):457–472, 1992.

[21] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *J. Res. Int. Educ.*, 5(2):189–211, 2012.

[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[23] M. D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res*, 15(1):1593–1623, 2014.

[24] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Found. Trends in Inf. Ret.*, 10(1):1–117, 2016.

[25] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

[26] J. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[27] B. Lambert. *A student's guide to Bayesian statistics*. Sage, 2018.

[28] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.

[29] C. Muth, Z. Oravecz, and J. Gabry. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quant. Methods Psychol.*, 14(2):99–119, 2018.

[30] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro. Using score distributions to compare statistical significance tests for information retrieval evaluation. *J. Am. Soc. Inf. Sci.*, 71(1):98–113, 2020.

[31] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In *Proc. SIGIR*, pages 5–14, 2016.

[32] T. Sakai. The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation. In *Proc. SIGIR*, pages 25–34, 2017.

[33] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.*, 11(5):447–470, 2008.

[34] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. SIGIR*, pages 623–632, 2007.

[35] D. X. Sousa, S. Canuto, M. A. Gonçalves, T. C. Rosa, and W. S. Martins. Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *ACM Trans. Inf. Sys.*, 37(2):1–34, 2019.

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

[37] E. M. Voorhees. Overview of TREC 2004 robust retrieval track. In *Proc. TREC*, pages 69–77, 2004.

[38] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proc. SIGIR*, pages 761–770, 2012.