# Sub-corpora Impact on System Effectiveness

Nicola Ferro
Department of Information Engineering,
University of Padua, Italy
ferro@dei.unipd.it

Mark Sanderson
Computer Science, School of Science, RMIT University,
Melbourne, Australia
mark.sanderson@rmit.edu.au

## ABSTRACT

Understanding the factors comprising IR system effectiveness is of primary importance to compare different IR systems. Effectiveness is traditionally broken down, using ANOVA, into a topic and a system effect but this leaves out a key component of our evaluation paradigm: the collections of documents. We break down effectiveness into topic, system and sub-corpus effects and compare it to the traditional break down, considering what happens when different evaluation measures come into play. We found that sub-corpora are a significant effect. The consideration of which allows us to be more accurate in estimating what systems are significantly different. We also found that the sub-corpora affect different evaluation measures in different ways and this may impact on what systems are considered significantly different.

## CCS CONCEPTS

**•Information systems →Evaluation of retrieval results; Test collections;**

## KEYWORDS

experimental evaluation; retrieval effectiveness; sub-corpus effect; effectiveness model; GLMM; ANOVA

## 1 INTRODUCTION

Studying the effectiveness of *Information Retrieval (IR)* systems is a core area of investigation, the main goal of which is to compare different IR systems in a robust and repeatable way. Commonly, IR system effectiveness is broken down as

$$\textit{effectiveness score} = \textit{topic effect} + \textit{system effect}$$

The topic effect was shown to be greater than the system effect using a two-way ANOVA to decompose effectiveness as above [1, 14]. The decomposition allowed simultaneous multiple comparisons of IR systems on TREC data, determining which were significantly better than others.

To improve the estimation of the system effect, you need to add components to the above model. For example, [10] showed that a topic*system interaction improved the estimation but the reported experiments relied on simulated data. Using a *Grid of Points (GoP)* approach (i.e. IR systems originated by a factorial combination of their components) the system effect can be sub-divided into system components in order to better understand system behavior [3].

However, at least one "ingredient" is missing from consideration: the *collections* of documents that are an integral part of the evaluation paradigm. Past work studied how sub-corpora impact IR effectiveness [13] and how collection size and the choice of documents influenced the way that a test collection ranked one retrieval system relative to another [7]. Both these studies highlighted the importance of sub-corpora to system performance but they did not incorporate the sub-corpus effect into a wider model:

$$\textit{effectiveness score} = \textit{topic effect} + \textit{system effect} + \textit{sub-corpus effect}$$

By integrating topic, system, and sub-corpus effects into the one model, comparisons can be made between the magnitude of the effects and, potentially, significant differences between systems can be more accurately calculated.

This paper addresses two research questions:

RQ1 what is the impact of considering sub-corpora in an effectiveness model?

RQ2 how do different evaluation measures behave with respect to effectiveness models including sub-corpus effects?

The methodology is described next (Sec. 2) followed by experiments and findings (Sec. 3), before finally concluding (Sec. 4).

## 2 METHODOLOGY

A *General Linear Mixed Model (GLMM)* [11] explains the variation of a dependent variable ("Data") in terms of a controlled variation of independent variables ("Model") in addition to a residual uncontrolled variation ("Error"): Data = Model + Error. In GLMM terms, *ANalysis Of VAriance (ANOVA)* attempts to explain data (dependent variable scores) in terms of the experimental conditions (the model) and an error component. Typically, ANOVA is used to determine under which condition dependent variables differ and what proportion of variation can be attributed to differences between specific conditions, as defined by the independent variable(s).

The experimental design determines how to compute the model and estimate its parameters. It is possible to have an *independent measures* design where different subjects participate in different experimental conditions (factors) or a *repeated measures* design, where each subject participates in all experimental conditions (factors). A final distinction is between *crossed/factorial* designs – where every level of one factor is measured in combination with every level of the other factors – and *nested* designs, where levels of a factor are grouped within each level of another nesting factor.

The traditional crossed repeated measures two-way ANOVA design, used in past work [1, 14], breaks down effectiveness into a topic and a system effect:

$$Y_{ij} = \mu_{\cdot\cdot} + \tau_i + \alpha_j + \varepsilon_{ij} \tag{1}$$

*SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan.*
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080674

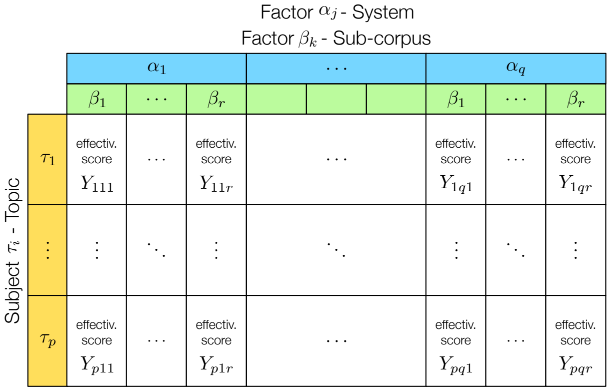

Factor $\alpha_j$ - System
Factor $\beta_k$ - Sub-corpus
$\alpha_1$
$\cdots$
$\alpha_q$
$\beta_1$ $\cdots$ $\beta_r$
$\beta_1$ $\cdots$ $\beta_r$
Subject $\tau_i$ - Topic
$\tau_1$
$\tau_p$
effectiv. score $Y_{111}$
effectiv. score $Y_{11r}$
effectiv. score $Y_{1q1}$
effectiv. score $Y_{1qr}$
effectiv. score $Y_{p11}$
effectiv. score $Y_{p1r}$
effectiv. score $Y_{pq1}$
effectiv. score $Y_{pqr}$

**Figure 1: Model for topic, system, and sub-corpus effects.**

where $Y_{ij}$ is the effectiveness score (from an evaluation measure) of the $i$-th subject in the $j$-th factor; $\mu_{\cdot\cdot}$ is the grand mean; $\tau_i = \mu_{i\cdot} - \mu_{\cdot\cdot}$ is the effect of the $i$-th subject, i.e. a topic, where $\mu_{i\cdot}$ is the mean of the $i$-th subject; $\alpha_j = \mu_{\cdot j} - \mu_{\cdot\cdot}$ is the effect of the $j$-th factor, i.e. a system, where $\mu_{\cdot j}$ is the mean of the $j$-th factor; finally, $\varepsilon_{ij}$ is the error committed by the model in predicting the effectiveness score of the $i$-th subject in the factor $j$. Examining eq (1) on both a whole and split collection (i.e. sub-corpora) we can understand changes to effectiveness between these two collection conditions.

We also explore a crossed repeated measures three-way ANOVA design, which breaks down effectiveness into a topic, system, and sub-corpus effect, as shown in Figure 1:

$$Y_{ijk} = \mu_{\cdot\cdot\cdot} + \tau_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \qquad (2)$$

where: $Y_{ijk}$ is the effectiveness score of the $i$-th subject in the $j$-th and $k$-th factors; $\mu_{\cdot\cdot\cdot}$ is the grand mean; $\tau_i = \mu_{i\cdot\cdot} - \mu_{\cdot\cdot\cdot}$ is the effect of the $i$-th subject, i.e. a topic, where $\mu_{i\cdot\cdot}$ is the mean of the $i$-th subject; $\alpha_j = \mu_{\cdot j\cdot} - \mu_{\cdot\cdot\cdot}$ is the effect of the $j$-th factor, i.e. a system, where $\mu_{\cdot j\cdot}$ is the mean of the $j$-th factor; $\beta_k = \mu_{\cdot\cdot k} - \mu_{\cdot\cdot\cdot}$ is the effect of the $k$-th factor, i.e. a sub-corpus, where $\mu_{\cdot\cdot k}$ is the mean of the $k$-th factor; $(\alpha\beta)_{jk}$ is the interaction between systems and sub-corpora; finally, $\varepsilon_{ijk}$ is the error committed by the model in predicting the effectiveness score of the $i$-th subject in the two factors $j$ and $k$.

We compare the GLMM models in eqs (1) and (2). Note, when we apply eq (1) to sub-corpora, we use the design shown in Figure 1 but omit the $\beta_k$ sub-corpus effect. Thus, we obtain a two-way ANOVA where we have more replicates for each (topic, system) pair, one for each sub-corpus.

An ANOVA test outcome indicates, for each factor, the *Sum of Squares (SS)*, the *Degrees of Freedom (DF)*, the *Mean Squares (MS)*, the F statistics, and the $p$-value of that factor, to determine significance. We are also interested in determining the proportion of variance that is due to a particular factor: i.e. we estimate its *effect-size measure* or *Strength of Association (SOA)*, which is a "*standardized index and estimates a parameter that is independent of sample size and quantifies the magnitude of the difference between populations or the relationship between explanatory and response variables*" [9].

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + N}$$

is an unbiased estimator of the variance components associated with the sources of variation in the design, where $F_{fact}$ is the F-statistic and $df_{fact}$ are the degrees of freedom for the factor while $N$ is the total number of samples. The common rule of thumb [11] when classifying $\hat{\omega}^2_{\langle fact \rangle}$ effect size is: $> 0.14$ is a *large size effect*, 0.06–0.14 is a *medium size effect*, and 0.01–0.06 is a *small size effect*. Negative $\hat{\omega}^2_{\langle fact \rangle}$ values are considered as zero.

In experimentation, a *Type I* error occurs if a true null hypothesis is rejected. The probability of such an error is $\alpha$. The chances of making a Type I error for a series of comparisons is greater than the error rate for a single comparison. If we consider $C$ comparisons, the probability of at least one Type I error is $1-(1-\alpha)^C$, which increases with the number of comparisons. Type I errors are controlled by applying the Tukey *Honestly Significant Difference (HSD)* test [5] with a significance level $\alpha = 0.05$. Tukey's method is used in ANOVA to create confidence intervals for all pairwise differences between factor levels, while controlling the family error rate. Two levels $u$ and $v$ of a factor are considered significantly different when

$$|t| = \frac{|\hat{\mu}_u - \hat{\mu}_v|}{\sqrt{MS_{error}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}} > \frac{1}{\sqrt{2}} q_{\alpha,k,N-k}$$

where $n_u$ and $n_v$ are the sizes of levels $u$ and $v$; $q_{\alpha,k,N-k}$ is the upper $100 * (1-\alpha)$th percentile of the studentized range distribution with parameter $k$ and $N-k$ degrees of freedom; $k$ is the number of levels in the factor and $N$ is the total number of observations.

## 3 EXPERIMENTS

We used the *TREC Adhoc* `T07` *and* `T08` collections: 528,155 documents made up of four `TIPSTER` sub-corpora: Foreign Broadcast Information Service (`TIPFBIS`, 130,471 documents); Federal Register (`TIPFR`, 55,630 documents); Financial Times (`TIPFT`, 210,158 documents); and Los Angeles Times (`TIPLA`, 131,896 documents). `T07` and `T08` provide 50 topics: `351-400` and `401-450`, as well as binary relevance judgments drawn from a pool depth of 100; 103 and 129 runs were submitted to `T07` and `T08`, respectively.

We split the `T07` and `T08` runs on the four sub-corpora by keeping the retrieved documents that belong to each sub-corpus. We applied the same split procedure to relevance judgments. This caused some topics to have no relevant documents on some sub-corpora, which suggests some kind of bias during topic creation and/or relevance assessment. Consequently, we kept only the topics that have at least one relevant document on each sub-corpus. This left us with 22 topics for `T07` and 15 topics for `T08`. We used eight evaluation measures: *Average Precision (AP)*, P@10; Rprec, *Rank-Biased Precision (RBP)* [8], *Normalized Discounted Cumulated Gain (nDCG)* [6], nDCG@20, *Expected Reciprocal Rank (ERR)* [2], and Twist [4].

Code to run the experiments is available at: https://bitbucket.org/frrncl/sigir2017-fs/.

### 3.1 RQ1 – Sub-corpora & effectiveness models

Figure 2 shows a worked example of the outcome of the application of the models on `T08` and AP. Figure 2(a) shows the ANOVA table for eq (1) on the whole collection. Both the topic and the system effects are significant and large: the system effect is about $\frac{3}{5}$ the

| Source | SS | DF | MS | F | p-value | $\hat{\omega}^2_{\langle fact \rangle}$ |
|---|---|---|---|---|---|---|
| **Topic** | 31.0056 | 14 | 2.2147 | 229.2833 | 0 | 0.6229 |
| **System** | 14.5575 | 128 | 0.1137 | 11.7744 | 5.774e-160 | 0.4161 |
| **Error** | 17.3092 | 1792 | 0.0097 | | | |
| **Total** | 62.8722 | 1934 | | | | |

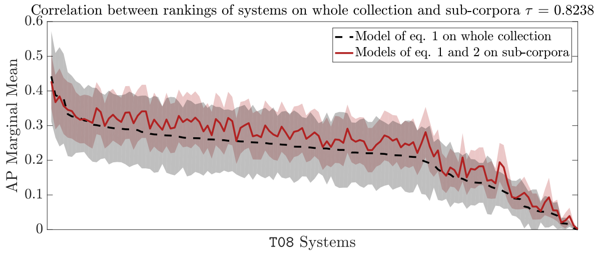
(a) ANOVA table for model of eq (1) on the whole collection.

| Source | SS | DF | MS | F | p-value | $\hat{\omega}^2_{\langle fact \rangle}$ |
|---|---|---|---|---|---|---|
| **Topic** | 181.1610 | 14 | 12.9401 | 326.3519 | 0 | 0.3705 |
| **System** | 62.2931 | 128 | 0.4867 | 12.2738 | 1.352e-220 | 0.1571 |
| **Error** | 301.2262 | 7597 | 0.0397 | | | |
| **Total** | 544.6802 | 7739 | | | | |

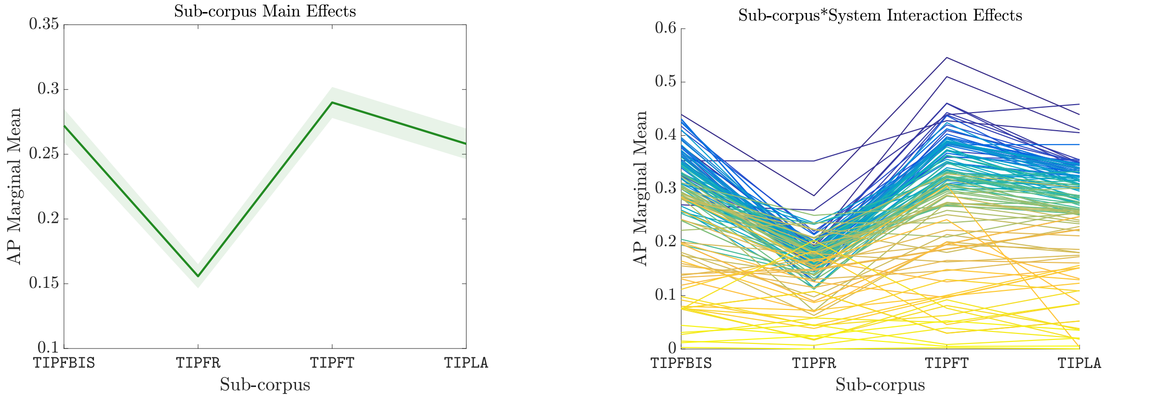
(b) ANOVA table for model of eq (1) on the sub-corpora.

| Source | SS | DF | MS | F | p-value | $\hat{\omega}^2_{\langle fact \rangle}$ |
|---|---|---|---|---|---|---|
| **Topic** | 181.1610 | 14 | 12.9401 | 349.9769 | 0 | 0.3870 |
| **System** | 62.2931 | 128 | 0.4867 | 13.1623 | 5.812e-238 | 0.1675 |
| **Sub-Corpus** | 21.0526 | 3 | 7.0175 | 189.7959 | 1.829e-118 | 0.0682 |
| **Sub-Corpus*System** | 13.5905 | 384 | 0.0354 | 0.9572 | 0.7137 | – |
| **Error** | 266.5831 | 7210 | 0.0370 | | | |
| **Total** | 544.6802 | 7739 | | | | |

(c) ANOVA table for model of eq (2) on the sub-corpora.



(d) Main effects plot for the system effect.



(e) Main effects plot for the sub-corpus effect.

(f) Interaction effects plot for the sub-corpus*system effect.

**Figure 2: Application of eq** (1) **and** (2) **to `T08` and AP both on the whole collection and the sub-corpora.**

size of the topic effect. These findings are consistent with past results [1, 14]. The Tukey HSD test detects 1,825 out of 8,256 (22.11%) possible system pairs as significantly different with 107 out of 129 (82.95%) systems being in the top-group, i.e. systems not significantly different from the top performing one. Figure 2(b) shows the model applied to the four sub-corpora. Both the topic and the system effects are significant and large, the system effect is about $\frac{2}{5}$ the size of the topic effect. The Tukey HSD test indicates that 1,872 out of 8,256 (22.67%) possible system pairs are significantly different with 64 out of 129 (49.61%) systems being in the top-group. Measuring on sub-corpora tends to decrease the size of the system effect relative to the topic effect. More pairs of significantly different systems were found with fewer in the top group.

Figure 2(d) plots the AP marginal mean of systems on the whole `TIPSTER` collection (black dashed line) and on the sub-corpora (red solid line) together with their confidence intervals (shaded). The AP values of systems change, but system ranking is not too dissimilar, as suggested by the Kendall's correlation $\tau = 0.8238$. We can see how the use of sub-corpora makes the confidence intervals smaller, suggesting more accuracy, as supported also by the outcomes of the Tukey HSD test.

Figure 2(c) shows eq (2) applied to the four sub-corpora. The SS of the topic and system effects is the same as in the case of Figure 2(b) but the SS of the error is reduced by the amount corresponding to the SS of the sub-corpus and sub-corpus*system effects. This makes the estimation of the effect size of the topic and system effects slightly more precise. The sub-corpus effect is a significant medium size effect, about $\frac{2}{5}$ of the system and $\frac{1}{5}$ of the topic effect, while the interaction between sub-corpora and systems is not significant. The Tukey HSD test reports that 1,993 out of 8,256 (24.14%) possible system pairs are significantly different with 71 out of 129 (55.04%) systems being in the top-group; this is coherent with the reduction of the $MS_{error}$ term which, being the other factors constant, makes the $|t|$ statistics in the Tukey HSD test bigger, thus detecting more significant differences.

Figure 2(e) shows the main effects plot for the sub-corpus effect: sub-corpora affect system effectiveness. Figure 2(f) plots the interaction effects for the sub-corpus*system effect where each line represents a different system. Even if, in the case of AP, the effect is not significant, we can note how sub-corpora affect systems differently. For example, the general trend is that systems have lower effectiveness on the `TIPFR` sub-corpus, even if a few systems behave the opposite way; similarly, `TIPFT` is the sub-corpus that results in highest effectiveness but with some exceptions.

| Track `T07` | AP | P@10 | R-prec | RBP | nDCG | nDCG@20 | ERR | Twist |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| $\hat{\omega}^2_{\langle \text{Topic} \rangle}$ | 0.4065 (0.00) | 0.2692 (0.00) | 0.3327 (0.00) | 0.2836 (0.00) | 0.4013 (0.00) | 0.3353 (0.00) | 0.2549 (0.00) | 0.3192 (0.00) |
| $\hat{\omega}^2_{\langle \text{System} \rangle}$ | 0.1639 (0.00) | 0.1050 (0.00) | 0.1319 (0.00) | 0.1151 (0.00) | 0.2625 (0.00) | 0.1624 (0.00) | 0.1155 (0.00) | 0.1500 (0.00) |
| $\hat{\omega}^2_{\langle \text{Sub-Corpus} \rangle}$ | 0.0075 (0.00) | 0.0838 (0.00) | 0.0181 (0.00) | 0.0878 (0.00) | 0.0048 (0.00) | 0.0087 (0.00) | 0.0844 (0.00) | 0.0207 (0.00) |
| $\hat{\omega}^2_{\langle \text{Sub-Corpus*System} \rangle}$ | – (0.43) | – (1.00) | – (0.53) | – (1.00) | 0.0230 (0.00) | 0.0112 (0.00) | – (0.87) | – (0.42) |
| $\tau$ | 0.9041 | 0.7746 | 0.8591 | 0.8062 | 0.8991 | 0.7164 | 0.7518 | 0.8386 |
| **Track `T08`** | AP | P@10 | R-prec | RBP | nDCG | nDCG@20 | ERR | Twist |
| $\hat{\omega}^2_{\langle \text{Topic} \rangle}$ | 0.3870 (0.00) | 0.2220 (0.00) | 0.2410 (0.00) | 0.2316 (0.00) | 0.4429 (0.00) | 0.4324 (0.00) | 0.2044 (0.00) | 0.2045 (0.00) |
| $\hat{\omega}^2_{\langle \text{System} \rangle}$ | 0.1675 (0.00) | 0.1162 (0.00) | 0.1232 (0.00) | 0.1335 (0.00) | 0.3207 (0.00) | 0.2135 (0.00) | 0.1417 (0.00) | 0.1515 (0.00) |
| $\hat{\omega}^2_{\langle \text{Sub-Corpus} \rangle}$ | 0.0682 (0.00) | 0.1310 (0.00) | 0.0650 (0.00) | 0.1631 (0.00) | 0.0491 (0.00) | 0.0498 (0.00) | 0.1710 (0.00) | 0.0964 (0.00) |
| $\hat{\omega}^2_{\langle \text{Sub-Corpus*System} \rangle}$ | – (0.71) | – (0.74) | – (0.75) | – (0.18) | 0.0141 (0.00) | – (0.22) | 0.0065 (0.04) | – (0.21) |
| $\tau$ | 0.8238 | 0.7229 | 0.7604 | 0.7682 | 0.8162 | 0.6696 | 0.6887 | 0.7772 |

**Table 1: Effect size ($\hat{\omega}^2$ SoA) and p-value for eq (2). Insignificant effects are in gray; small effects, light blue; medium, blue; and large, dark blue. The $\tau$ reports system ranking correlation when using the whole collection and sub-corpora.**

### 3.2 RQ2 – Sub-corpora & evaluation measures

Table 1 shows eq (2) applied to the four sub-corpora for `T07` and `T08` for all evaluation measures. The topic effect is significant and large in all cases while the system effect is a significant medium size effect in about half of the cases and large in the other half.

The sub-corpora are always a significant effect with small or medium size, except for RBP and ERR on `T08` for which it is a large size. On `T07`, the sub-corpus effect is always smaller than the system effect, on `T08` the sub-corpus effect is bigger than the system effect for P@10, RBP, and ERR. The sub-corpus*system interaction effect is generally not significant, with the exception of nDCG and nDCG@20 on `T07` and nDCG and ERR on `T08` for which it is significant though small.

Table 1 shows the Kendall's $\tau$ correlations between the rankings of systems using eq (1) on the whole `TIPSTER` collection and eq (2) on the four sub-corpora. The rankings are generally correlated, indicating a good agreement between the two approaches, even if there are some cases where correlation drops, namely P@10, nDCG@20, and ERR, on `T08` and nDCG@20 on `T07`.

## 4 CONCLUSION AND FUTURE WORK

We find that sub-corpora are a significant effect on system effectiveness. While past work has indicated such an effect exists, to the best of our knowledge, this is the first time such an effect has been integrated into a effectiveness model and effect sizes compared to other known factors. We find that different evaluation measures are affected in different ways by sub-corpora, which may impact on what systems are considered significantly different to each other. We found that ranking systems using sub-corpora reasonably agrees with ranking systems with respect to a whole collection but using the information about sub-corpora allows a more accurate estimation of which systems are significantly different.

This is initial work. We recognize that the number of topics in our collections is small. We next plan to understand the impact of different kinds of sub-corpora. We also plan to extend the present methodology to study the impact of different collections on system performance rather than sub-corpora within one collection.

## 5 ACKNOWLEDGMENTS

This work is supported in part by the Australian Research Council's Discovery Projects scheme (DP130104007) and a Google Faculty Award.

## REFERENCES

[1] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval* 1, 1-2, 7–34.

[2] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *CIKM 2009*. 621–630.

[3] N. Ferro and G. Silvello. 2016. A General Linear Mixed Models Approach to Study System Component Effects. In *SIGIR 2016*. 25–34.

[4] N. Ferro, G. Silvello, H. Keskustalo, A. Pirkola, and K. Järvelin. 2016. The Twist Measure for IR Evaluation: Taking User's Effort Into Account. *JASIST* 67, 3, 620–648.

[5] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. John Wiley & Sons, USA.

[6] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM TOIS* 20, 4, 422–446.

[7] T. Jones, A. Turpin, S. Mizzaro, F. Scholer, and M. Sanderson. 2014. Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation. In *CIKM 2014*. 1843–1846.

[8] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS* 27, 1, 2:1–2:27.

[9] S. Olejnik and J. Algina. 2003. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods* 8, 4, 434–447.

[10] S. E. Robertson and E. Kanoulas. 2012. On Per-topic Variance in IR Evaluation. In *SIGIR 2012*. 891–900.

[11] A. Rutherford. 2011. *ANOVA and ANCOVA. A GLM Approach* (2nd ed.). John Wiley & Sons, New York, USA.

[12] T. Sakai. 2014. Statistical Reform in Information Retrieval? *SIGIR Forum* 48, 1, 3–12.

[13] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. 2012. Differences in Effectiveness Across Sub-collections. In *CIKM 2012*. 1965–1969.

[14] J. M. Tague-Sutcliffe and J. Blustein. 1994. A Statistical Analysis of the TREC-3 Data. In *TREC-3*. 385–398.