On the Measurement of Test Collection Reliability

Julián Urbano[*]
jurbano@inf.uc3m.es

Mónica Marrero[*]
mmarrero@inf.uc3m.es

Diego Martín[†]
dmartin@dit.upm.es

[*]University Carlos III of Madrid
Department of Computer Science
Leganés, Spain

[†]Technical University of Madrid
Department of Telematics Engineering
Madrid, Spain

ABSTRACT

The reliability of a test collection is proportional to the number of queries it contains. But building a collection with many queries is expensive, so researchers have to find a balance between reliability and cost. Previous work on the measurement of test collection reliability relied on data-based approaches that contemplated random *what if* scenarios, and provided indicators such as swap rates and Kendall tau correlations. Generalizability Theory was proposed as an alternative founded on analysis of variance that provides reliability indicators based on statistical theory. However, these reliability indicators are hard to interpret in practice, because they do not correspond to well known indicators like Kendall tau correlation. We empirically established these relationships based on data from over 40 TREC collections, thus filling the gap in the practical interpretation of Generalizability Theory. We also review the computation of these indicators, and show that they are extremely dependent on the sample of systems and queries used, so much that the required number of queries to achieve a certain level of reliability can vary in orders of magnitude. We discuss the computation of confidence intervals for these statistics, providing a much more reliable tool to measure test collection reliability. Reflecting upon all these results, we review a wealth of TREC test collections, arguing that they are possibly not as reliable as generally accepted and that the common choice of 50 queries is insufficient even for stable rankings.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation.*

General Terms

Experimentation, Measurement, Reliability.

Keywords

Test Collection, Evaluation, Reliability, Generalizability Theory, TREC.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

1. INTRODUCTION

The purpose of evaluating the effectiveness of an Information Retrieval (IR) system is to assess how well it would satisfy real users. The main tool used in these evaluations are test collections, which comprise a collection of documents to search, a set of queries $\mathcal{Q}$, and a set of relevance judgments that contains information as to what documents are relevant, and to which degree, to the queries [16]. Given the results returned by a system A for one of the queries $q \in \mathcal{Q}$, an effectiveness measure uses the information in the relevance judgments to compute a score $\lambda_{q,\mathsf{A}}$ that represents the effectiveness of the system for that query. After running the system for all queries in the collection, the average $\overline{\lambda}_{\mathcal{Q},\mathsf{A}} = \frac{1}{|\mathcal{Q}|}\sum \lambda_{i,\mathsf{A}}$ is usually reported as the main measure of system effectiveness, representing the expected behavior of the system for an arbitrary new query. When comparing two systems A and B, the main measure reported is the average effectiveness difference $\overline{\Delta\lambda}_{\mathcal{Q},\mathsf{AB}} = \overline{\lambda}_{\mathcal{Q},\mathsf{A}} - \overline{\lambda}_{\mathcal{Q},\mathsf{B}}$. Based on this difference, we conclude which system is better.

The immediate question to ask is: *how reliable are those conclusions about system effectiveness?* Ideally, researchers would evaluate the system with the set of all possible queries that a user might request. In such a case, we could be sure that the true average performance of the system corresponds to the score we computed with the collection. The problem is that building such a collection is either impractical for requiring an enormous amount of queries and relevance judgments, or just impossible if the potential query set is not defined, which use to be the case because we can not account for future queries that do not yet exist. Therefore, the query set $\mathcal{Q}$ in a test collection must be regarded as a sample from the universe of all queries, and the sample mean $\overline{\lambda}_{\mathcal{Q},\mathsf{A}}$ as an estimate of the *true* effectiveness mean λ_{A}. But because we are estimating this score with a sample of queries, our estimates are erroneous to some degree. The results may change drastically with a different query set $\mathcal{Q}'$, so much that differences between systems could be reversed.

An evaluation result is reliable if it can be replicated with another collection: if the set of queries $\mathcal{Q}$ suggests that system A outperforms system B, we can be very sure that the conclusion would hold for a different set of queries $\mathcal{Q}'$, and in the end, for the universe of all queries. A simple way to make a collection reliable is to include many queries; the more we employ the smaller the variance of the estimates and thus the more reliable the conclusion. The problem is that more queries also means more cost to create the collection, so researchers have to find a balance between the reliability of the results and the cost of the collection. To this end, it is necessary to develop indicators of test collection reliability.

Several works in the last fifteen years have studied the problem of reliability in IR evaluation experiments. The basic methodology consisted in evaluating a series of systems with two different and random sets of queries, computing several reliability indicators that measured how similar those evaluations were. Using different query sample sizes and randomizing query selection, researchers were able to map query set size to reliability and extrapolate results to larger query sets. The data used consisted in runs submitted to several TREC tracks (mostly the Ad Hoc tracks), and the sets of queries employed in each edition. While these approaches are clearly faithful to the data, they are limited in that the full query set had to be partitioned in two disjoint sets to comply with the assumption that they were independent.

In 2007 Bodoff and Li [6] proposed Generalizability Theory (GT) as an alternative [7, 18]. GT is grounded on analysis of variance components, which allows to dissect the variability in effectiveness scores and figure out how much of it is due to system differences, query difficulty, assessors, etc. In an ideal evaluation setting, we would like all variance to be due to actual differences between systems and not due to query variability; if the queries in the collection are too varied, or differences between systems too small, then we need many queries to ensure that our estimates are reliable. From these variance components GT allows researchers to estimate the reliability of a test collection even before it has been created. Based on some previous data, GT can estimate the reliability of a collection with a larger number of queries, more than one assessor providing judgments for each query, etc. GT provides indicators for the stability of both the absolute scores and the relative differences by computing different variance ratios.

The main advantages of GT against the traditional data-based approaches are that 1) it is based on statistical theory, 2) it is easy to employ because it does not require tedious and repetitive *what if* scenarios, and 3) it allows us to estimate the reliability of a collection or experimental design that does not exist yet. But it has disadvantages too: 1) it is unknown the extent to which reliability indicators are affected by the data used to estimate variance components, and 2) it is very hard to interpret them in practical terms.

We address these two problems of GT applied to the measurement of test collection reliability. In the next section we review past work following data-based approaches and the reliability indicators used. We then review the use of GT and discuss the motivation for this work. In Section 3 we show how the initial data used in GT studies has a very large effect on the results, discussing minimum sample sizes and interval estimators. Section 4 reports a study to provide an empirical mapping between GT-based indicators of reliability and the well known data-based ones. Next we discuss the reliability of several TREC collections based on the results from previous sections, presenting conclusions in Section 6.

2. INDICATORS OF RELIABILITY

Several indicators of test collection reliability have been proposed in the literature. This section reviews traditional indicators found in the early data-based studies and the GT-based indicators more recently proposed.

2.1 Data-based Indicators

Given a query set $\mathcal{Q}$ and a similar set $\mathcal{Q}'$ of the same size, we can define the following data-based reliability indicators:

- **Kendall correlation ($\boldsymbol{\tau}$),** compares the order in which systems are ranked according to $\mathcal{Q}$ and $\mathcal{Q}'$, regardless of the magnitude of the differences $\overline{\Delta\lambda}_{\mathsf{AB}}$. It ranges from 1 (same rankings) to -1 (reversed rankings), counting the number of system pairs that are swapped between the two rankings. For $\mathcal{Q}$ to be reliable, τ must therefore tend to 1.
- **AP correlation ($\boldsymbol{\tau_{AP}}$),** adds a top-heaviness component to Kendall τ, such that swaps between systems towards the top of the rankings are more penalized than swaps towards the bottom [23].
- **Power ratio ($\boldsymbol{\beta}$),** is the fraction of pairwise system differences that result statistically significant according to query set $\mathcal{Q}$. If the difference $\overline{\Delta\lambda}_{\mathcal{Q},\mathsf{AB}}$ between two systems is deemed as statistically significant, it serves as further evidence that the true difference $\Delta\lambda_{\mathsf{AB}}$ has the same sign. For $\mathcal{Q}$ to be reliable, β must therefore tend to 100%. In this paper we compute standard 2-tailed t-tests at the 0.05 level [19].
- **Minor Conflict ratio ($\boldsymbol{\alpha_-}$),** is the fraction of statistically significant differences with $\mathcal{Q}$ that have a sign swap with $\mathcal{Q}'$ but are not statistically significant there. α_- is therefore the fraction of *uncertain* conclusions when measuring statistical significance, so for $\mathcal{Q}$ to be reliable α_- must therefore tend to 0%.
- **Major Conflict ratio ($\boldsymbol{\alpha_+}$),** is the fraction of statistically significant differences with $\mathcal{Q}$ that are also significant with $\mathcal{Q}'$ but have a sign swap. α_+ is therefore the fraction of *incorrect* conclusions when measuring statistical significance, so for $\mathcal{Q}$ to be reliable α_+ must therefore tend to 0% as well.
- **Absolute Sensitivity ($\boldsymbol{\delta_a}$),** is the minimum *absolute* difference $\overline{\Delta\lambda}_{\mathcal{Q},\mathsf{AB}}$ that need be observed between any two systems such that the differences with $\mathcal{Q}'$ have the same sign at least 95% of the times. For $\mathcal{Q}$ to be reliable, δ_a must therefore tend to 0, meaning that even small differences can be trusted.
- **Relative Sensitivity ($\boldsymbol{\delta_r}$),** is the minimum *relative* difference $\overline{\Delta\lambda}_{\mathcal{Q},\mathsf{AB}}/\max\left(\overline{\lambda}_{\mathcal{Q},\mathsf{A}},\overline{\lambda}_{\mathcal{Q},\mathsf{B}}\right)$ that need be observed with $\mathcal{Q}$ such that the differences with $\mathcal{Q}'$ have the same sign at least 95% of the times. For $\mathcal{Q}$ to be reliable, δ_r must therefore tend to 0% too.
- **Root Mean Squared Error ($\boldsymbol{\varepsilon}$),** measures the difference between the absolute scores with $\mathcal{Q}$ and with $\mathcal{Q}'$. Thus, for $\mathcal{Q}$ to be reliable ε must tend to 0 too.

One of the first reliability studies was conducted in 1998 by Voorhees [20], who analyzed the effect of having different assessors provide relevance judgments. Employing a methodology based on randomization, she concluded that the absolute scores could suffer wide variations between assessors, but that the ranking of systems was seldom altered, establishing $\tau = 0.9$ as the de facto minimum on ranking similarity. She also studied swap rates as a function of $\overline{\Delta\lambda}$ and suggested a minimum of 25 queries to have a somewhat stable ranking. Also in 1998, Zobel [24] studied the effect of pool depth on absolute system scores, extrapolating trends to larger pool depths. He also compared different statistical procedures in terms of power and conflict ratios.

Buckley and Voorhees [8] compared in 2000 the reliability of various effectiveness measures by mapping effectiveness differences to error rates. Extrapolating to 50 queries, they concluded that $\overline{\Delta\lambda} \geq 0.05$ produced less than 1.5% system swaps when computing Average Precision (AP), while other measures such as Precision at cutoff 10 (P@10) produced 3.6% of swaps. In 2002, Voorhees and Buckley [22] extended their work with other collections and methods, but again extrapolating trends. They concluded that with 50 queries the sensitivity of AP was $\delta_a = 0.05$, while increasing the query set size to 100 would yield $\delta_a = 0.03$. They also reported large differences across collections and effectiveness measures. Lin and Hauptmann [13] showed that the empirical model used by Voorhees and Buckley can be derived theoretically, and that the three factors affecting reliability are query set size, mean effectiveness scores, and variability of scores. Sanderson and Zobel [17] also revisited this work by computing relative sensitivity and incorporating statistical procedures to account for score variability. They concluded $\delta_r = 10\%$ with AP if coupled with statistical significance, and $\delta_r = 25\%$ if not. They observed very similar relative sensitivity between AP and P@10, arguing the use of more queries with fewer judgments as previous work suggested that much of the score variability is due to queries [4].

In 2007 Sakai [15] used similar methods to compare the reliability of several effectiveness measures, though he did not extrapolate to larger query sets. He computed τ correlations, absolute sensitivity δ_a and a variation of δ_r, and observed that these indicators were not very correlated with statistical significance, arguing the importance of considering score variability rather than just means. Voorhees revisited in 2009 [21] the use of statistical procedures with the TREC Robust 2004 collection, computing reliability indicators with an unprecedented set of 100 queries, therefore avoiding the need to extrapolate to the usual size 50. When using AP, she observed power $\beta = 47\%$ and conflict ratios $\alpha_- = 2.7\%$ and $\alpha_+ = 0.04\%$. She showed again that P@10 is less reliable than AP also in these terms; and that nDCG showed higher reliability (agreeing with Sakai [15]). She also found that minor conflicts were usually coupled with large relative differences, thus suggesting that researchers employ several large collections to draw general conclusions.

2.2 GT-based Indicators

Bodoff and Li [6] proposed Generalizability Theory [7, 18] as an alternative to measure test collection reliability that directly addresses variability of scores rather than just the mean as was common before. GT has two stages: a Generalizability study (G-study) to estimate variance components based on previous data, and a Decision study (D-study) that subsequently computes reliability indicators for a different experimental design. We consider a fully crossed design and decompose variability of scores into three components: variance due to actual differences among systems (σ_s^2), variance due to differences in difficulty among queries (σ_q^2), and variance due to the system-query interaction effect whereby some systems are particularly good (or bad) for some queries ($\sigma_{s:q}^2$). The variance due to other effects, such as assessors, is in our case confounded with the interaction effect.

Using Analysis of Variance (ANOVA) procedures, these variance components can be estimated from previous data:

$$\hat{\sigma}_{s:q}^2 = \hat{\sigma}_e^2 = \mathrm{E}M_{residual} \tag{1}$$

$$\hat{\sigma}_s^2 = \frac{\mathrm{E}M_s - \hat{\sigma}_e^2}{n_q} \tag{2}$$

$$\hat{\sigma}_q^2 = \frac{\mathrm{E}M_q - \hat{\sigma}_e^2}{n_s} \tag{3}$$

where $\mathrm{E}M_\nu$ is the expected Mean Square of component ν, and n_s and n_q are the number of systems and queries [7, 18]. These estimates can be used to compute the proportion of total variance that is due to each of the effects, such as how much of it is due to actual differences between systems.

In the D-study, we can use the variance estimates from the G-study to compute the reliability of a larger query set. To this end, two reliability indicators are usually employed:

- **Generalizability Coefficient ($\mathbf{E\rho^2}$),** is the ratio of system variance to itself plus relative error variance:

$$\mathrm{E}\rho^2\left(n'_q\right) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_e^2}{n'_q}} \tag{4}$$

  and it provides a measure of the stability of relative differences between systems $\overline{\Delta\lambda}$. By extension, it measures the reliability of the ranking. For a collection to be reliable, $\mathrm{E}\rho^2$ must therefore tend to 1.

- **Index of Dependability ($\mathbf{\Phi}$),** is the ratio of system variance to itself plus absolute error variance:

$$\Phi\left(n'_q\right) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_q^2 + \sigma_e^2}{n'_q}} \tag{5}$$

  and it provides a measure of the stability of absolute effectiveness scores $\overline{\lambda}$. For a collection to be reliable, Φ must therefore tend to 1 as well.

The main advantage of these indicators is that they allow us to estimate the reliability of an arbitrary query set size n'_q, so there is no need to follow the traditional methodologies based on random *what if* scenarios and extrapolation. From equations (4) and (5) it can be seen that the reliability of the collection increases as n'_q increases, because the estimates of query difficulty (i.e. average system performance per query) are more precise. These indicators were used by Kanoulas and Aslam [12] to derive the gain and discount functions of nDCG that yield optimal reliability when n'_q is constant.

With simple algebraic manipulation, we can calculate the minimum number of queries needed to reach some level of relative or absolute stability π:

$$n'_{\mathrm{E}\rho^2}(\pi) = \left\lceil \frac{\pi \cdot \sigma_e^2}{\sigma_s^2 (1-\pi)} \right\rceil \tag{6}$$

$$n'_{\Phi}(\pi) = \left\lceil \frac{\pi\left(\sigma_q^2 + \sigma_e^2\right)}{\sigma_s^2 (1-\pi)} \right\rceil \tag{7}$$

which can be used to estimate how many more queries we need to add to our collection for it to be reliable. The main use of this approach can be found in the TREC Million Query Track [2, 1], which set out to study whether many queries with a few judgments yield more reliable results than a few queries with many judgments. The conclusion was that $n'_q \approx 80$ queries are sufficient for a reliable ranking, while $n'_q \approx 130$ are needed for reliable absolute scores.

2.3 Motivation

The two problems of GT can be clearly spotted at this point. First, equations (1) to (3) show that we do not compute the *true* σ_ν^2 variance components, but just *estimates* $\hat{\sigma}_\nu^2$ based on some previous data. If we use a different, yet similar set of systems or queries to estimate these variance components, the resulting $\mathrm{E}\hat{\rho}^2$ and $\hat{\Phi}$ scores might be very

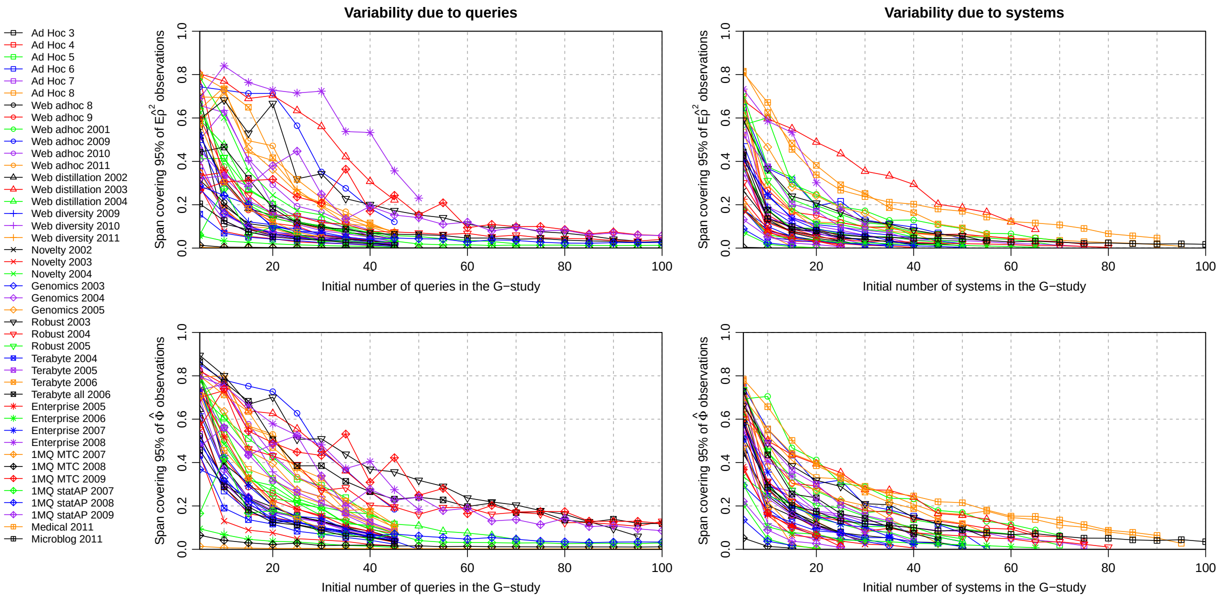

Figure 1: Variability in E$\hat{\rho}^2$ (top) and $\hat{\Phi}$ (bottom) scores as a function of the initial number of queries (left) and number of systems (right) used in the G-study to estimate variance components.

different. In a revised paper, Bodoff [5, §4.6] briefly discussed this issue and argued that differences are marginal. However, he reports the results when randomly selecting only one system per research group instead of all of them, and only one trial of such experiment. We argue that this situation is not representative because the full set of systems and the reduced set after removing runs by the same groups are actually very similar to begin with, so it is expected that reliability scores do not change much. Also, only one such randomly reduced set is compared, so there is really no evidence to support that claim. Likewise, he further suggests that as few as five queries or systems are often enough to provide stable estimates of the variance components in the G-study [5, §3.1]. We further analyze this issue in Section 3.

Second, equations (6) and (7) allow us to estimate the minimum number of queries n'_q to reach some stability level π, but the greater question is: *how much is stable enough?* Bodoff [5] mentions that in most Social Science applications a stability coefficient of 0.8 is acceptable, but there is no similar standard for Engineering applications. Kanoulas and Aslam [12] set $\Phi = 0.95$ as the target in their experiments, but this choice is arbitrary. In their analysis of the Million Query Track 2007 [2] and 2008 [1], Allan et al. [1] and Carterette et al. [9, 10] also set E$\rho^2 = 0.95$ as the target. They mention in a footnote that in their experiments E$\rho^2 = 0.95$ approximately corresponded to $\tau = 0.9$, but details are omitted. We study this issue in Section 4 by empirically mapping GT-based indicators onto data-based indicators that are easier to understand and use in practice.

3. VARIABILITY OF GT INDICATORS

To measure the effect of the number of queries and number of systems used in the G-study to estimate variance components, we use data from 43 TREC collections covering 12 tasks across 10 tracks, from TREC 3 to TREC 2011 (see Table 1). As in previous studies [22, 17, 6, 21], we remove the bottom 25% of systems so that our results are not obscured by possibly buggy implementations. For each collection, we randomly selected $n_q = 5$ queries and computed the variance components using the full set of systems. We then calculated Eρ^2 and Φ for the full query set size, and the required number of queries to reach 0.95 stability. This was repeated with increments in n_q of 5 queries, up the maximum permitted by the collection or 100. For each query set size, we ran 200 random trials, each of which can be considered as the possible data available for a G-study when analyzing a test collection design. The same process was followed by varying the initial number of systems n_s and using the full set of queries instead.

Figure 1 shows the variability in G-study results[1]. For each collection and initial number of queries used, the y-axis plots the length of the span covering 95% of the E$\hat{\rho}^2$ and $\hat{\Phi}$ observations in the 200 random trials. The right hand side plots show the same span lengths, but for different number of systems used in the G-study. As expected, the queries have a larger effect. Most importantly, we see that the average span length with just 5 queries is about 0.5 across collections. That is, the stability estimates could be as low as 0.3 or as high 0.8, for example, just depending on the particular set of queries we use in the G-study. In fact, estimates of the minimum number of queries required can vary in orders of magnitude if not using enough data. For example, with as many as 30 initial queries and all 184 systems from the Microblog 2011 collection, GT may suggest from 63 to 133 queries to reach E$\rho^2 = 0.95$. Similarly, from 40 initial systems and all 34 queries from the Medical 2011 collection, GT may suggest from 109 to 566 queries. In general, at least 50 queries and 50 systems seem necessary for 95% of estimates to be within a 0.1 span. This means that GT may be trusted to measure the reliability of an existing collection, but that

[1] Given the amount of datapoints displayed in this paper, we recommend to access the full-color version available online.

$$\left(\frac{L_\zeta - 1}{n_q}, \frac{U_\zeta - 1}{n_q}\right), \text{where} \qquad (8)$$

$$L_\zeta = \frac{M_s}{M_e F_{\alpha:df_s,df_e}}$$

$$U_\zeta = \frac{M_s}{M_e F_{1-\alpha:df_s,df_e}}$$

$$\left(\frac{n_s L_\Lambda}{n_s L_\Lambda + n_q}, \frac{n_s U_\Lambda}{n_s U_\Lambda + n_q}\right), \text{where} \qquad (9)$$

$$L_\Lambda = \frac{M_s^2 - F_{\alpha:df_s,\infty} M_s M_e + \left(F_{\alpha:df_s,\infty} - F_{\alpha:df_s,df_e}\right) F_{\alpha,df_s,df_e} M_e^2}{(n_s - 1) F_{\alpha:df_s,\infty} M_s M_e + F_{\alpha:df_s,df_q} M_s M_q}$$

$$U_\Lambda = \frac{M_s^2 - F_{1-\alpha:df_s,\infty} M_s M_e + \left(F_{1-\alpha:df_s,\infty} - F_{1-\alpha:df_s,df_e}\right) F_{1-\alpha,df_s,df_e} M_e^2}{(n_s - 1) F_{1-\alpha:df_s,\infty} M_s M_e + F_{1-\alpha:df_s,df_q} M_s M_q}$$

researchers should be cautious when planning a collection based on the results of a handful of systems and queries.

These results clearly evidence the need for a measure of confidence on GT indicators. Bodoff [5] suggests the use of confidence intervals to account for this variability, but only computes them for the variance components in the G-study. Confidence intervals for the ultimately more useful D-study can be worked out from various variance ratios (see equations (8) and (9)[2]). Feldt [11] derived exact $100(1-2\alpha)\%$ confidence intervals for the ratio $\zeta = \sigma_s^2/\sigma_e^2$ under the assumption of normally distributed scores. The confidence interval on $\mathrm{E}\rho^2(n'_q)$ is computed using the endpoints in (8):

$$\mathrm{E}\rho^2\left(n'_q\right) = \frac{n'_q \zeta}{1 + n'_q \zeta} \qquad (10)$$

Arteaga et al. [3] derived approximate $100(1-2\alpha)\%$ confidence intervals for the ratio $\Lambda = \sigma_s^2/\left(\sigma_s^2 + \sigma_q^2 + \sigma_e^2\right)$, again assuming a normal distribution of scores. The confidence interval on $\Phi\left(n'_q\right)$ is computed using the endpoints in (9):

$$\Phi\left(n'_q\right) = \frac{n'_q \Lambda}{1 + \left(n'_q - 1\right)\Lambda} \qquad (11)$$

Brennan [7, §6] discusses different methods to compute confidence intervals in both G-studies and D-studies, showing that the above intervals work reasonably well even when the normality assumption is violated. The right hand side of Table 1 reports the point and 95% interval estimates of the stability of the 43 TREC collections we consider in this paper. These intervals provide a more suitable estimate of test collection reliability because they account for variability in the G-study. For example, researchers could use these intervals to infer the required number of queries to reach the lower endpoint of the interval instead of the point estimate:

$$n'_{\mathrm{E}\rho^2}(\pi) = \left\lceil \frac{\pi}{\zeta(1-\pi)} \right\rceil \qquad (12)$$

$$n'_{\Phi}(\pi) = \left\lceil \frac{\pi(1-\Lambda)}{\Lambda(1-\pi)} \right\rceil \qquad (13)$$

4. INTERPRETING GT INDICATORS

To empirically derive a mapping between GT-based and data-based reliability indicators, we again used the 43 TREC collections in Table 1. For each collection we proceeded as follows. Two random and disjoint query subsets of size $n_q = 10$ were selected from the full set of queries; let these subsets be $\mathcal{Q}$ and $\mathcal{Q}'$. The full set of systems was evaluated with both query subsets, and all data-based reliability indicators in Section 2.1 were computed, along with the two GT-based indicators according to $\mathcal{Q}$ and $\mathcal{Q}'$. This was repeated with increments in n_q of 10 queries, up to the maximum permitted by the collection. For query subset size we ran 50 random trials, each trial providing us with 32 datapoints ($\mathrm{E}\hat{\rho}^2$ and $\hat{\Phi}$ according to $\mathcal{Q}$ and to $\mathcal{Q}'$, mapped to $\hat{\tau}$,$\hat{\tau}_{AP}$, $\hat{\beta}$, $\hat{\alpha}_-$, $\hat{\alpha}_+$, $\hat{\delta}_a$, $\hat{\delta}_r$ and $\hat{\varepsilon}$). Theoretically though, $\mathrm{E}\rho^2$ is better related to τ, τ_{AP}, β, α_-, α_+ and δ_a because it measures the stability of relative differences, while Φ is better related to δ_r and ε because it measures the stability of absolute scores. We thus mapped only these combinations.

Figure 2 shows the mappings. For each collection we fitted a model with all available datapoints. However, we dropped points for which $\mathrm{E}\hat{\rho}^2 < 0.8$ and $\hat{\Phi} < 0.5$ so that the trends were not affected by mappings with such small stability to be even practical. These thresholds were chosen based on the observed stability of the 43 TREC collections; about 85% of them show larger stability scores (see Table 1). This resulted in over 28,000 points for each plot. In the top three plots (τ, τ_{AP} and β) we fitted the model $y = x^a$, where a is the parameter to fit. This resulted in the desired theoretical behavior that $\lim_{x\to 1} y = 1$ and $\lim_{x\to 0} y = 0$, that is, when all variability is due to system differences τ should be 1 because the ranking cannot be altered, and if all variance is due to queries then τ should be 0 because the rankings are completely random. Similarly, in the bottom four plots we fitted the model $y = (1-x)^a$, such that $\lim_{x\to 1} y = 0$ and $\lim_{x\to 0} y = 1$, that is, ε should for example be 0 if there is no variability due to queries.

As the first plot shows, all 43 collections do actually need $\mathrm{E}\rho^2 > 0.95$ to reach $\tau = 0.9$. In general, $\mathrm{E}\rho^2 = 0.95$ corresponds to $\tau \approx 0.85$, and on average $\mathrm{E}\rho^2 \approx 0.97$ is needed across collections to reach $\tau = 0.9$. The two clear exceptions are found in the Million Query Track. The 2008 collection is the one that reaches the target $\tau = 0.9$ with the lowest stability ($\mathrm{E}\rho^2 \approx 0.93$), while the 2007 collection needs the largest ($\mathrm{E}\rho^2 \approx 0.98$). Note that these were the two collections for which the $\mathrm{E}\rho^2 = 0.95 \to \tau = 0.9$ correspondence was established [1, 9, 10]. It should be noted here that these fits have an exponential-like shape, meaning that it is hard to achieve a mid level of τ, but once $\mathrm{E}\rho^2$ is large enough small improvements in stability translate into large improvements in τ. However, the relation between n'_q and $\mathrm{E}\rho^2$ has a logarithmic-like shape, meaning that it is increasingly more expensive to improve $\mathrm{E}\rho^2$ to begin with. Thus, it should be considered the required effort for slight improvements in τ.

The second plot shows quite high τ_{AP} scores at these levels of relative stability, but generally below τ. This suggests that the swaps in the rankings are still happening between systems at the top of the rankings [23]. The third plot shows that at these stability levels it is expected to observe statistical significance in about 80% of system comparisons. In the middle right plot we can see that the proportion of conflicting results is generally below the $\alpha = 0.05$ significance level when $\mathrm{E}\rho^2 \geq 0.9$.

[2] $F_{\varphi:df_1,df_2}$ is the quantile function of the F distribution with df_1 and df_2 degrees of freedom. In our fully crossed design, $df_s = n_s - 1$, $df_q = n_q - 1$, and $df_e = (n_s-1)(n_q-1)$.

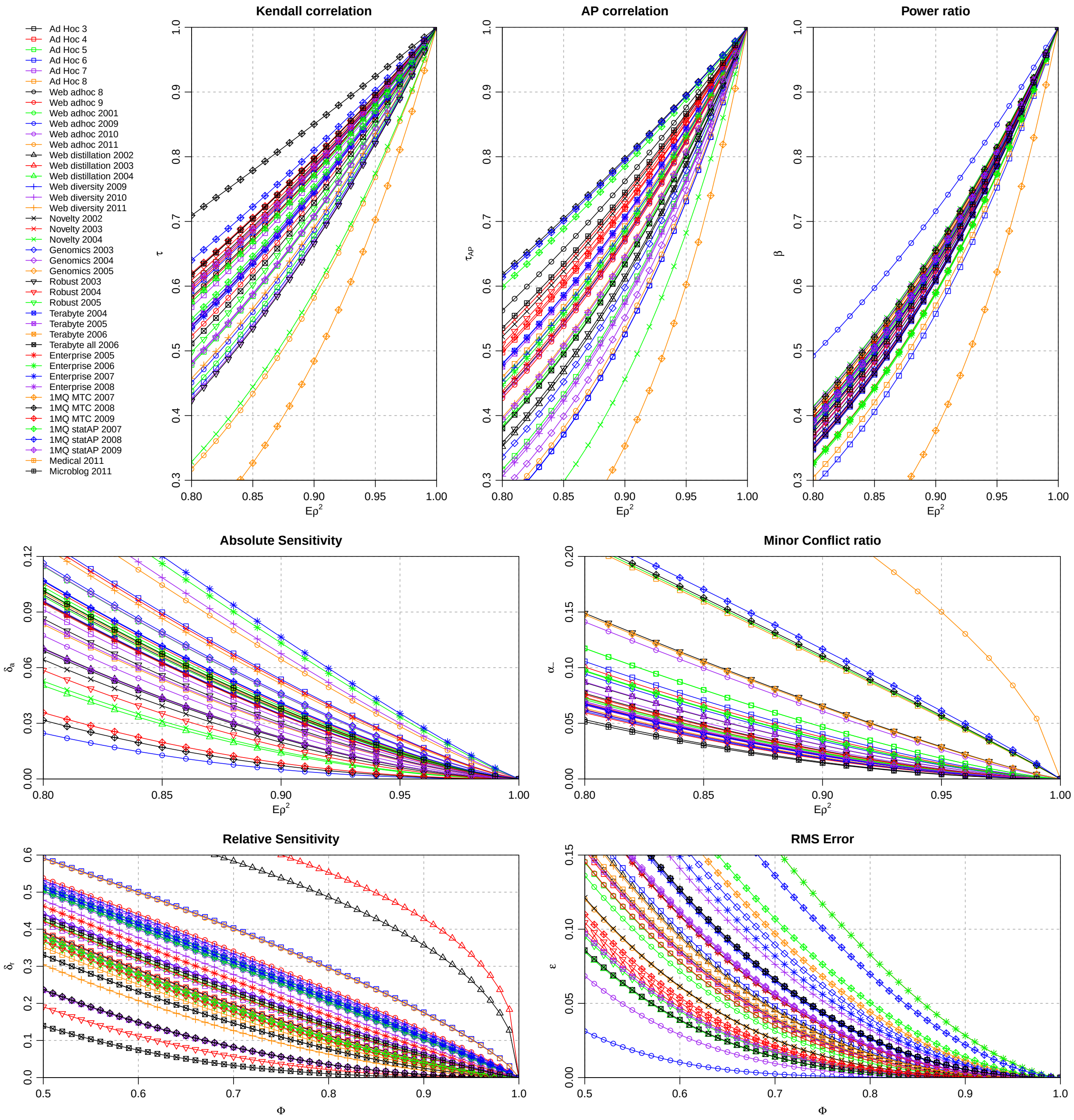

Figure 2: Mapping from GT-based to data-based reliability indicators on a per-collection basis.

Researchers interested in the particular mapping for one of these collections may use the estimates in Table 1 and the plots in Figure 2 to get a better understanding of the evaluation results and draw more informed conclusions. To assess the reliability of future collections and guide in their development process, we fitted a single model using all available data instead of one model per collection. Figure 3 shows these fits, along with 95% and 90% *prediction* intervals that theoretically cover 95% and 90% of all future observations. In terms of sensitivity, the middle left plots show that $\delta_a \approx 0.03$ for $\mathrm{E}\rho^2 \approx 0.9$, which is about 60% of what Voorhees and Buckley reported for the Ad Hoc tracks [22]; although the intervals cover their values well. In the bottom left plot we see that $\delta_r \approx 20\%$ for $\Phi \approx 0.75$, generally agreeing with Sanderson and Zobel [17]. As to statistical significance, we replicated Voorhees's [21] study with random sets of 50 queries from the Ad Hoc 7-8 topics and Robust 2004 systems. The average relative stability is $\mathrm{E}\hat{\rho}^2 \in [0.81, 0.88]$, which corresponds to $\beta \in [37\%, 54\%]$, $\alpha_- \in [3.9\%, 7.8\%]$ and $\alpha_+ \in [0.38\%, 1.3\%]$. These are again larger than she reported, but the intervals cover her values well.

Overall, these models produce a decent fit on the data, and they fill the gap between data-based methodologies and Generalizability Theory. They provide a valuable tool to rapidly assess and easily understand the reliability of a test collection design.

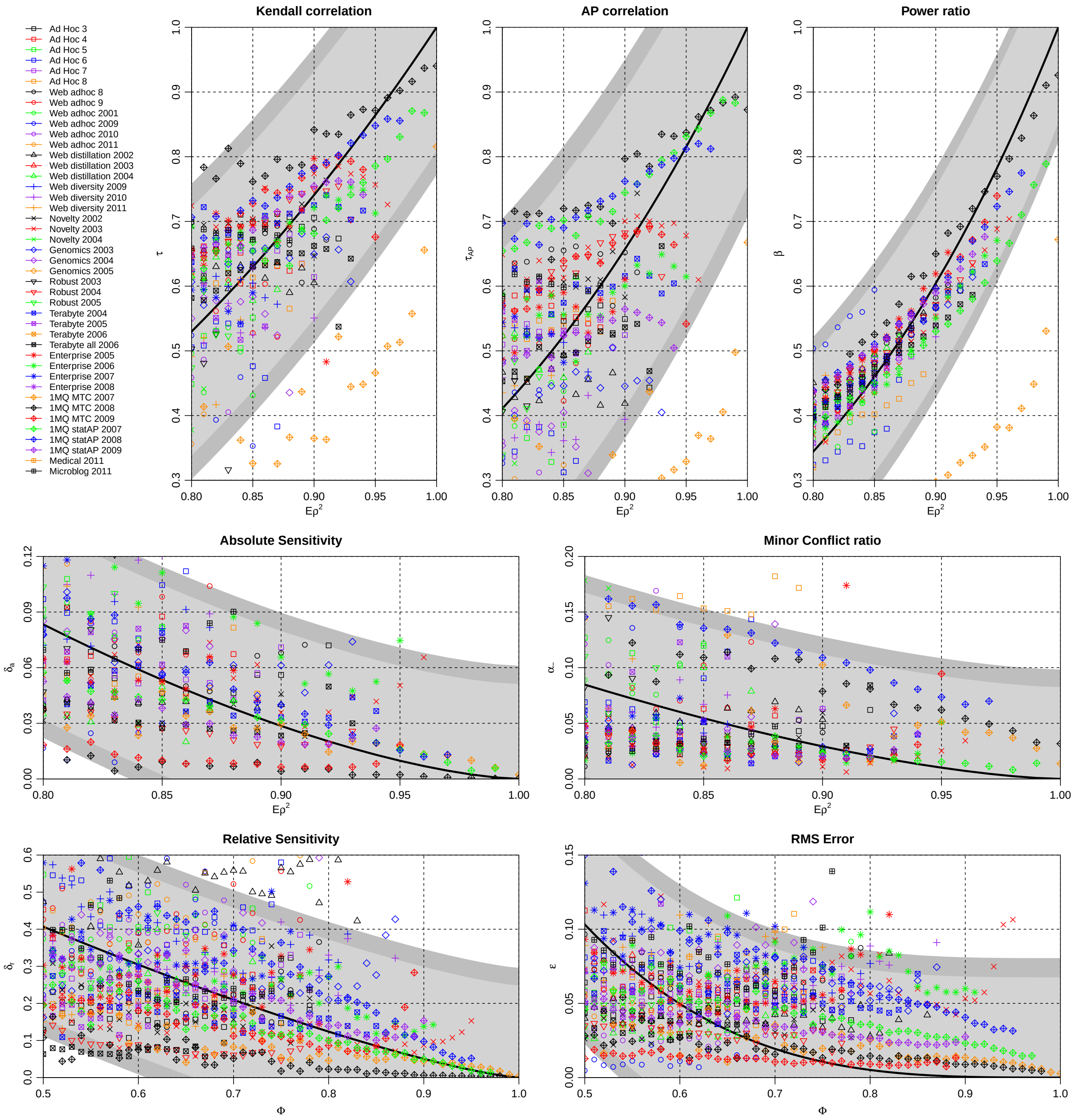

Figure 3: General mapping from GT-based to data-based reliability indicators, with 95% (dark grey) and 90% (light grey) prediction intervals.

5. DISCUSSION

The last columns in Table 1 report point and 95% interval estimates of the stability of the 43 TREC collections we considered. Collections in the same group correspond to the same tasks, providing a historical perspective on the reliability of the collections used so far since 1994 and for a variety of tasks. For example, the average relative stability in the Ad Hoc collections was $E\rho^2 \in [0.86, 0.93]$, which according to Figure 3 corresponds to $\tau \in [0.65, 0.81]$. For the Web Ad Hoc collections we find $E\rho^2 \in [0.8, 0.93]$, which would correspond to $\tau \in [0.53, 0.81]$. There are large differences within some tasks, such as Web Distillation, Genomics, Terabyte and Enterprise. This is further evidence of the variability in D-study results due to the data used in the G-study. Except for a few particular cases though, the computation of confidence intervals smooths the problem. Across collections the averages are $E\rho^2 = 0.88$ and $\Phi = 0.74$, with some tasks having very low scores. According to Figure 3 the expected τ correlation is 0.69 with variations from 0.49 to 0.95, that is, much lower than desired.

Figure 4 plots the historical trend of test collection reliability. The left plot shows that relative stability has varied in the (0.8,1) interval for the most part, but most importantly it suggests that the stability of collections has decreased very

Track	Documents	Query Set	Measure	n_s	n_q	$\mathrm{E}\hat{\rho}^2(n_q)$		$\hat{\Phi}(n_q)$	
Ad Hoc 3	Disks 1 & 2	151-200	AP	40	50	0.933	0.893-0.963	0.786	0.661-0.88
Ad Hoc 4	Disks 2 & 3	201-250	AP	33	49	0.907	0.847-0.952	0.79	0.658-0.89
Ad Hoc 5	Disks 2 & 4	251-300	AP	94	50	0.856	0.804-0.9	0.62	0.488-0.732
Ad Hoc 6	Disks 4 & 5	301-350 *	AP	74	50	0.898	0.855-0.933	0.806	0.714-0.875
Ad Hoc 7	Disks 4 & 5	351-400 *	AP	103	50	0.919	0.891-0.943	0.799	0.71-0.864
Ad Hoc 8	Disks 4 & 5	401-450 *	AP	129	50	0.908	0.88-0.932	0.701	0.59-0.787
Web$_{\text{AdHoc}}$ 8	WT2g	401-450 *	AP	44	50	0.929	0.89-0.96	0.83	0.728-0.904
Web$_{\text{AdHoc}}$ 9	WT10g	451-500	AP	104	50	0.876	0.833-0.912	0.76	0.662-0.835
Web$_{\text{AdHoc}}$ 2001	WT10g	501-550	AP	97	50	0.862	0.813-0.904	0.711	0.598-0.801
Web$_{\text{AdHoc}}$ 2009	ClueWeb09	“W1-W50” *	AP (MTC)	71	50	0.81	0.729-0.876	0.619	0.473-0.744
Web$_{\text{AdHoc}}$ 2010	ClueWeb09	“W51-W100” *	AP	56	48	0.829	0.746-0.895	0.662	0.513-0.787
Web$_{\text{AdHoc}}$ 2011	ClueWeb09	“W101-W150” *	AP	37	50	0.804	0.685-0.895	0.702	0.537-0.835
Web$_{\text{Distillation}}$ 2002	.GOV	551-600	AP	71	49	0.901	0.858-0.935	0.84	0.762-0.898
Web$_{\text{Distillation}}$ 2003	.GOV	TD1-TD50	AP	93	50	0.45	0.249-0.619	0.315	0.144-0.492
Web$_{\text{Distillation}}$ 2004	.GOV	“WT04”	AP	74	75	0.89	0.844-0.927	0.747	0.643-0.832
Web$_{\text{Diversity}}$ 2009	ClueWeb09	“W1-W50” *	α-nDCG@20	48	50	0.903	0.852-0.943	0.847	0.759-0.911
Web$_{\text{Diversity}}$ 2010	ClueWeb09	“W51-W100” *	α-nDCG@20	32	50	0.882	0.803-0.94	0.804	0.676-0.899
Web$_{\text{Diversity}}$ 2011	ClueWeb09	“W101-W150” *	α-nDCG@20	25	50	0.844	0.725-0.929	0.719	0.535-0.865
Novelty 2002	Disks 4 & 5	50 from 300-450 *	F	42	49	0.919	0.873-0.955	0.792	0.671-0.883
Novelty 2003	AQUAINT	N1-N50	F	55	50	0.966	0.949-0.979	0.944	0.91-0.967
Novelty 2004	AQUAINT	N51-N100	F	60	50	0.801	0.708-0.876	0.181	0.1-0.301
Genomics$_{\text{AdHoc}}$ 2003	MEDLINE	“G1-G50”	AP	49	50	0.94	0.909-0.965	0.87	0.792-0.925
Genomics$_{\text{AdHoc}}$ 2004	MEDLINE	“G51-G100”	AP	43	50	0.903	0.848-0.945	0.768	0.64-0.868
Genomics$_{\text{AdHoc}}$ 2005	MEDLINE	“G101-150”	AP	62	49	0.77	0.664-0.855	0.422	0.269-0.586
Robust 2003	Disks 4 & 5	50 from 301-450 & 601-650 *	AP	78	100	0.846	0.784-0.897	0.509	0.384-0.636
Robust 2004	Disks 4 & 5	301-450 & 601-700 *	AP	110	249	0.95	0.934-0.964	0.824	0.768-0.872
Robust 2005	AQUAINT	50 from 301-700 *	AP	74	50	0.864	0.807-0.911	0.693	0.564-0.797
Terabyte 2004	GOV2	701-750 *	bpref	70	49	0.953	0.933-0.97	0.877	0.809-0.924
Terabyte 2005	GOV2	751-800 *	bpref	58	50	0.875	0.815-0.923	0.648	0.501-0.774
Terabyte 2006	GOV2	801-850 *	bpref	80	50	0.762	0.668-0.841	0.427	0.283-0.575
Terabyte$_{\text{All}}$ 2006	GOV2	701-850 *	bpref	61	149	0.94	0.913-0.962	0.719	0.617-0.812
Enterprise$_{\text{Expert}}$ 2005	W3C	EX01-EX50	AP	37	50	0.916	0.864-0.955	0.824	0.713-0.905
Enterprise$_{\text{Expert}}$ 2006	W3C	EX51-EX105	AP	91	49	0.965	0.952-0.976	0.939	0.909-0.96
Enterprise$_{\text{Expert}}$ 2007	CERC	CE001-CE050	AP	55	50	0.884	0.827-0.929	0.785	0.674-0.87
Enterprise$_{\text{Expert}}$ 2008	CERC	CE051-CE127	AP	42	55	0.565	0.315-0.757	0.28	0.11-0.498
1MQ 2007	GOV2	“MQ1-MQ10000”	AP (MTC)	29	1692	0.999	0.999-1	0.998	0.997-0.999
1MQ 2008	GOV2	“MQ10001-MQ20000”	AP (MTC)	25	784	0.998	0.996-0.999	0.988	0.979-0.995
1MQ 2009	ClueWeb09	“MQ20001-MQ60000”	AP (MTC)	35	542	0.96	0.936-0.979	0.908	0.854-0.951
1MQ 2007	GOV2	“MQ1-MQ10000”	statAP	29	1153	0.992	0.986-0.996	0.982	0.97-0.991
1MQ 2008	GOV2	“MQ10001-MQ20000”	statAP	25	564	0.978	0.962-0.99	0.969	0.946-0.986
1MQ 2009	ClueWeb09	“MQ20001-MQ60000”	statAP	35	475	0.96	0.935-0.979	0.929	0.886-0.963
Medical 2011	NLP	“M101-M135”	bpref	127	34	0.774	0.704-0.835	0.497	0.348-0.628
Microblog 2011	Tweets2011	MB1-MB50	P@30	184	49	0.92	0.899-0.938	0.818	0.747-0.869

Table 1: Summary of all 43 TREC collections analyzed. Query sets with * are used in more than one collection. Query numbers in quotes are not official, but arbitrarily named for this paper. The last two columns report the point and 95% interval estimates of the GT-based reliability indicators.

slightly with the years. The clear exceptions are again the Million Query Track collections, which specifically aimed at increasing the number of queries. Within each task it appears that stability tended to decrease as the tasks got older despite that query set sizes were normally unaltered. The second plot shows that this decrease in stability could be due to system variance getting smaller with the years. That is, systems perform more similarly as the tasks get older, indicating that retrieval techniques are generally improved. The right plot shows that query difficulty also varied within tasks. Sudden peaks may be explained by changes in the document set or in the task definition. The general trend suggests that queries are getting more alike with the years, further contributing to the decrease in reliability.

Bodoff [5, §5] discusses the incorporation of the document set as another facet in Generalizability Theory, much like queries and systems, to measure variability due to documents [14]. He argues that it does not make sense in general, because we do no assign performance scores for individual documents but for sets of documents (e.g. the first k retrieved when computing $P@k$). In our case we could compare different editions of the same task but with different document sets to get a (weak) clue of the variability due to documents. For example, the Ad Hoc task of the Web Track shows quite different stability scores in the first three editions (WT2g and WT10g collections) compared to the last three editions (ClueWeb09), given that they all used the standard query set size of 50. Similarly, the Expert Search task in the Enterprise Track shows very different stability levels when using the W3C collection or the CERC collection. We must bear in mind though that these differences might actually be due to the systems and queries used, which varied from year to year.

From the confidence intervals in Table 1, we used the models fitted in Section 4 to provide in Table 2 the estimated data-based reliability scores for all 43 collections. It is evident that expected τ correlations are well below the desired 0.9 in most cases. In that line, some collections are clearly

Track	$\hat{\tau}$	$\hat{\tau}_{AP}$	$\hat{\beta}$ (%)	$\hat{\alpha}_-$ (%)	$\hat{\alpha}_+$ (%)	$\hat{\delta}_a$	$\hat{\delta}_r$ (%)	$\hat{\epsilon}$	$\hat{n}'_{E\rho^2}(.95)$	$\hat{n}'_{\Phi}(.95)$
Ad Hoc 3	0.725-0.898	0.637-0.86	58-83	0.6-3.2	0.02-0.28	0.01-0.03	6-25	0.001-0.029	37-114	130-487
Ad Hoc 4	0.622-0.87	0.515-0.823	45-79	0.9-5.6	0.03-0.72	0.01-0.06	6-25	0.001-0.03	47-169	116-484
Ad Hoc 5	0.537-0.741	0.418-0.657	35-60	2.9-8.2	0.23-1.38	0.03-0.08	18-42	0.013-0.112	106-233	348-999
Ad Hoc 6	0.641-0.821	0.537-0.758	47-72	1.6-5.2	0.08-0.62	0.02-0.05	7-20	0.001-0.017	69-161	136-381
Ad Hoc 7	0.72-0.846	0.631-0.791	58-76	1.2-3.3	0.05-0.29	0.01-0.03	8-20	0.001-0.017	58-117	150-389
Ad Hoc 8	0.695-0.819	0.6-0.756	54-72	1.6-3.9	0.08-0.38	0.02-0.04	13-31	0.006-0.054	69-130	257-662
$\text{Web}_{\text{AdHoc}}$ 8	0.718-0.89	0.629-0.849	57-82	0.7-3.4	0.02-0.3	0.01-0.03	5-18	0-0.014	40-118	102-355
$\text{Web}_{\text{AdHoc}}$ 9	0.595-0.77	0.484-0.694	42-65	2.4-6.4	0.17-0.9	0.02-0.06	10-24	0.003-0.028	92-190	189-484
$\text{Web}_{\text{AdHoc}}$ 2001	0.554-0.749	0.437-0.668	37-62	2.8-7.7	0.21-1.22	0.03-0.08	12-31	0.005-0.051	102-220	236-640
$\text{Web}_{\text{AdHoc}}$ 2009	0.406-0.686	0.283-0.59	22-53	4.1-13.5	0.41-3.24	0.04-0.13	17-44	0.011-0.122	135-354	327-1058
$\text{Web}_{\text{AdHoc}}$ 2010	0.434-0.729	0.311-0.643	25-59	3.1-12.2	0.27-2.73	0.03-0.12	13-39	0.006-0.095	107-311	247-868
$\text{Web}_{\text{AdHoc}}$ 2011	0.34-0.728	0.221-0.642	16-59	3.2-17	0.27-4.81	0.03-0.17	10-37	0.003-0.08	112-438	188-819
$\text{Web}_{\text{Distillation}}$ 2002	0.647-0.827	0.544-0.766	48-73	1.5-5	0.07-0.59	0.01-0.05	5-16	0.001-0.009	65-154	106-292
$\text{Web}_{\text{Distillation}}$ 2003	0.019-0.255	0.004-0.148	0-10	22.8-64.4	7.89-47.06	0.23-0.64	41-82	0.108-0.6	585-2862	980-5631
$\text{Web}_{\text{Distillation}}$ 2004	0.617-0.807	0.508-0.741	44-70	1.8-5.8	0.1-0.76	0.02-0.06	10-26	0.003-0.034	112-264	288-791
$\text{Web}_{\text{Diversity}}$ 2009	0.633-0.847	0.528-0.792	46-76	1.2-5.4	0.05-0.66	0.01-0.05	4-16	0-0.009	58-166	93-301
$\text{Web}_{\text{Diversity}}$ 2010	0.535-0.839	0.416-0.782	35-74	1.3-8.3	0.06-1.4	0.01-0.08	5-23	0.001-0.025	61-234	107-457
$\text{Web}_{\text{Diversity}}$ 2011	0.401-0.811	0.278-0.746	22-70	1.7-13.8	0.09-3.35	0.02-0.14	7-37	0.001-0.081	73-360	149-826
Novelty 2002	0.679-0.877	0.582-0.833	52-80	0.9-4.2	0.03-0.44	0.01-0.04	6-24	0.001-0.026	44-136	124-457
Novelty 2003	0.86-0.941	0.81-0.919	78-90	0.3-1.1	0-0.04	0-0.01	1-4	0-0	21-52	33-94
Novelty 2004	0.374-0.685	0.252-0.589	19-53	4.1-15.1	0.42-3.93	0.04-0.15	63-87	0.309-0.709	135-392	2203-8579
$\text{Genomics}_{\text{AdHoc}}$ 2003	0.762-0.903	0.684-0.867	63-84	0.6-2.5	0.02-0.18	0.01-0.02	3-13	0-0.006	35-95	78-250
$\text{Genomics}_{\text{AdHoc}}$ 2004	0.624-0.852	0.517-0.799	45-76	1.2-5.6	0.05-0.71	0.01-0.05	7-27	0.001-0.035	56-171	146-536
$\text{Genomics}_{\text{AdHoc}}$ 2005	0.311-0.641	0.195-0.537	14-47	5.2-18.8	0.62-5.69	0.05-0.19	32-67	0.055-0.358	158-472	657-2528
Robust 2003	0.5-0.734	0.379-0.649	31-60	3.1-9.6	0.25-1.78	0.03-0.09	27-53	0.036-0.204	218-525	1087-3043
Robust 2004	0.823-0.902	0.761-0.865	72-84	0.6-1.6	0.02-0.08	0.01-0.02	7-15	0.001-0.008	175-336	693-1428
Robust 2005	0.544-0.766	0.426-0.689	36-64	2.5-8	0.17-1.32	0.02-0.08	13-34	0.005-0.066	94-227	242-733
Terabyte 2004	0.82-0.916	0.758-0.884	72-86	0.5-1.6	0.01-0.08	0-0.02	4-12	0-0.004	30-68	77-220
Terabyte 2005	0.558-0.795	0.442-0.725	38-68	2-7.5	0.12-1.19	0.02-0.07	15-41	0.008-0.103	80-217	279-947
Terabyte 2006	0.316-0.61	0.2-0.5	14-44	6-18.5	0.8-5.52	0.06-0.18	33-65	0.06-0.336	181-474	702-2406
$\text{Terabyte}_{\text{All}}$ 2006	0.772-0.897	0.696-0.859	65-83	0.7-2.4	0.02-0.16	0.01-0.02	11-29	0.004-0.043	111-269	657-1761
$\text{Enterprise}_{\text{Expert}}$ 2005	0.661-0.877	0.56-0.831	50-80	0.9-4.7	0.03-0.52	0.01-0.05	5-20	0-0.017	46-149	100-383
$\text{Enterprise}_{\text{Expert}}$ 2006	0.868-0.932	0.821-0.907	79-89	0.3-1	0.01-0.03	0-0.01	2-4	0-0	24-48	39-93
$\text{Enterprise}_{\text{Expert}}$ 2007	0.582-0.812	0.468-0.746	40-70	1.7-6.8	0.09-1	0.02-0.07	7-23	0.001-0.025	73-200	143-459
$\text{Enterprise}_{\text{Expert}}$ 2008	0.037-0.453	0.01-0.33	0-26	11.4-56	2.41-37.02	0.11-0.56	41-86	0.104-0.683	335-2277	1053-8458
1MQ 2007	0.997-0.999	0.995-0.999	99-100	0-0	0-0	0-0	0-0	0-0	11-38	30-104
1MQ 2008	0.989-0.997	0.985-0.996	98-100	0-0	0-0	0-0	0-1	0-0	16-59	81-313
1MQ 2009	0.827-0.942	0.767-0.92	73-90	0.3-1.5	0-0.07	0-0.01	2-8	0-0.002	219-710	534-1756
1MQ 2007	0.962-0.989	0.947-0.984	94-98	0-0.1	0-0	0-0	0-1	0-0	88-304	196-685
1MQ 2008	0.896-0.972	0.858-0.961	83-95	0.1-0.7	0-0.02	0-0.01	0-2	0-0	107-421	156-616
1MQ 2009	0.826-0.941	0.765-0.919	73-90	0.3-1.5	0-0.08	0-0.01	1-6	0-0.001	194-628	352-1156
Medical 2011	0.368-0.598	0.246-0.486	19-42	6.3-15.5	0.88-4.08	0.06-0.15	28-57	0.039-0.246	129-273	383-1208
Microblog 2011	0.74-0.833	0.656-0.774	60-74	1.4-3	0.07-0.24	0.01-0.03	7-17	0.001-0.011	62-105	141-315

Table 2: Predicted reliability of all 43 TREC collections analyzed. All confidence intervals are based on the fits from Figure 3 at the endpoints of the 95% confidence intervals computed with equations (10) and (11).

not reliable, such as the Web Distillation 2003, Genomics Ad Hoc 2005, Terabyte 2006, Enterprise Expert Search 2008, or the very recent Medical 2011 and Web Ad Hoc 2011. Regarding the expected RMS Error of absolute scores, we can see that collections are somewhat stable, but with clear exceptions such as Web Distillation 2003, Novelty 2004 and Enterprise Expert Search 2008.

The last two columns in Table 2 report intervals on the number of queries, as per equations (12) and (13), required to achieve 0.95 stability. In general the number of queries needs to be at least doubled, and in many cases a few hundred queries seem to be needed. This is particularly interesting for the most recent collections, such as Web Ad Hoc 2010 and 2011, Medical 2011 and Microblog 2011, which stick to the traditional size of 50 queries but need about 200. What becomes clear from these figures is that the ideal size of a collection depends greatly on the task it will be used for, and thus it is not appropriate to fix some acceptable size such as 50 or 100 throughout tasks. Each task has different characteristics and should be analyzed accordingly.

6. CONCLUSIONS

In this paper we discussed the measurement of test collection reliability from the perspective of traditional data-based methodologies and of Generalizability Theory. GT is regarded as a more appropriate, easy to use, and powerful method to assess reliability, but it has two drawbacks. First, we showed that GT is very sensitive to the particular sample of systems and queries used to estimate reliability of a larger query set. We showed that about 50 systems and 50 queries are needed for robust estimates of collection reliability. Therefore, researchers should be cautious in using GT when building new collections from scratch. To account for all this variability we discussed a more robust approach based on interval estimates of the stability indicators, which helps in making more appropriate decisions regarding number of queries or different structure in the experimental design. Second, we empirically established a mapping between GT-based and traditional data-based indicators to help interpreting results from GT which, otherwise, do not have a

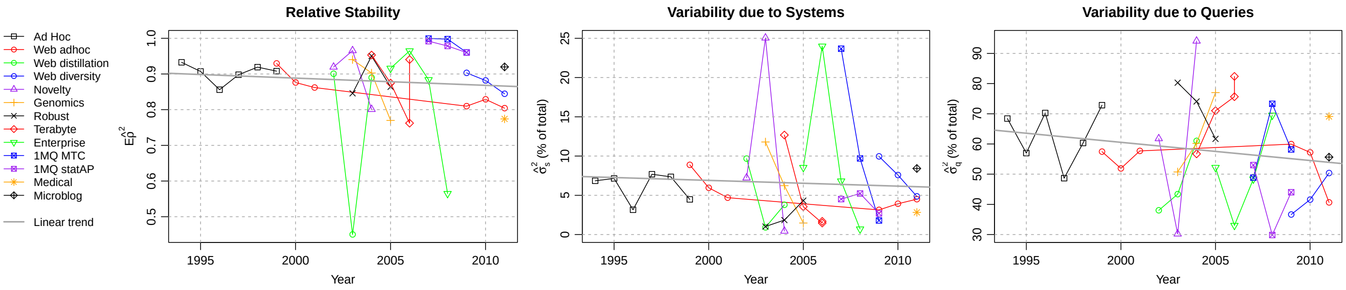

Figure 4: Historical trend of relative stability (left), variability due to systems (middle) and to queries (right).

clear and easily understandable meaning. Based on these results, we reviewed the reliability of 43 TREC test collections, evidencing that some of them are very little reliable. We show that the traditional choice of 50 queries is clearly not enough even for stable rankings, and in most cases a couple hundred queries are needed. Our results also show that the ideal query set size varies significantly across tasks, suggesting that we avoid the use of some fixed size such as 50 or 100 and that we analyze tasks and collections separately.

There are two clear lines for future research. First, we completely ignored the assessor facet in our study. It is evident that different assessors provide different results, so it would be interesting to include them in the analysis. Second, although we fitted the theoretically correct models, it is clear that they can be improved (see for instance Power and RMS Error in Figure 3). IR evaluation experiments generally violate assumptions of GT, such as normality of distributions and random sampling, so different models and features to better fit the actual data should be investigated.

We created some scripts for the statistical software `R` that can help researchers perform all these computations to easily assess the reliability of custom test collection designs. They can be downloaded from `http://julian-urbano.info`.

7. REFERENCES

[1] J. Allan, J. A. Aslam, B. Carterette, V. Pavlu, and E. Kanoulas. Million Query Track 2008 Overview. In *Text REtrieval Conference*, 2008.

[2] J. Allan, B. Carterette, J. A. Aslam, V. Pavlu, B. Dachev, and E. Kanoulas. Million Query Track 2007 Overview. In *Text REtrieval Conference*, 2007.

[3] C. Arteaga, S. Jeyaratnam, and G. A. Franklin. Confidence Intervals for Proportions of Total Variance in the Two-Way Cross Component of Variance Model. *Communications in Statistics: Theory and Methods*, 11(15):1643–1658, 1982.

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

[5] D. Bodoff. Test Theory for Evaluating Reliability of IR Test Collections. *Information Processing and Management*, 44(3):1117–1145, 2008.

[6] D. Bodoff and P. Li. Test Theory for Assessing IR Test Collections. In *ACM SIGIR*, pages 367–374, 2007.

[7] R. L. Brennan. *Generalizability Theory*. Springer, 2001.

[8] C. Buckley and E. M. Voorhees. Evaluating Evaluation Measure Stability. In *ACM SIGIR*, pages 33–34, 2000.

[9] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation Over Thousands of Queries. In *ACM SIGIR*, pages 651–658, 2008.

[10] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. If I Had a Million Queries. In *ECIR*, pages 288–300, 2009.

[11] L. S. Feldt. The Approximate Sampling Distribution of Kuder-Richardson Reliability Coefficient Twenty. *Psychometrika*, 30(3):357–370, 1965.

[12] E. Kanoulas and J. A. Aslam. Empirical Justification of the Gain and Discount Function for nDCG. In *ACM CIKM*, pages 611–620, 2009.

[13] W.-H. Lin and A. Hauptmann. Revisiting the Effect of Topic Set Size on Retrieval Error. In *ACM SIGIR*, pages 637–638, 2005.

[14] S. Robertson and E. Kanoulas. On Per-Topic Variance in IR Evaluation. In *ACM SIGIR*, pages 891–900, 2012.

[15] T. Sakai. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. *Information Processing and Management*, 43(2):531–548, 2007.

[16] M. Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[17] M. Sanderson and J. Zobel. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *ACM SIGIR*, pages 162–169, 2005.

[18] R. J. Shavelson and N. M. Webb. *Generalizability Theory: A Primer*. Sage Publications, 1991.

[19] J. Urbano, M. Marrero, and D. Martín. A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation. In *ACM SIGIR*, 2013.

[20] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *ACM SIGIR*, pages 315–323, 1998.

[21] E. M. Voorhees. Topic Set Size Redux. In *ACM SIGIR*, pages 806–807, 2009.

[22] E. M. Voorhees and C. Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *ACM SIGIR*, pages 316–323, 2002.

[23] E. Yilmaz, J. A. Aslam, and S. Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *ACM SIGIR*, pages 587–594, 2008.

[24] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? In *ACM SIGIR*, pages 307–314, 1998.