When Measurement Misleads: The Limits of Batch Assessment of Retrieval Systems

Justin Zobel
The University of Melbourne
Parkville, Victoria, Australia
jzobel@unimelb.edu.au

Abstract

The discipline of information retrieval (IR) has a long history of examination of how best to measure performance. In particular, there is an extensive literature on the practice of assessing retrieval systems using batch experiments based on collections and relevance judgements. However, this literature has only rarely considered an underlying principle: that measured scores are inherently incomplete as a representation of human activity, that is, there is an innate gap between measured scores and the desired goal of human satisfaction. There are separate challenges such as poor experimental practices or the shortcomings of specific measures, but the issue considered here is more fundamental – straightforwardly, in batch experiments the human-machine gap cannot be closed. In other disciplines, the issue of the gap is well recognised and has been the subject of observations that provide valuable perspectives on the behaviour and effects of measures and the ways in which they can lead to unintended consequences, notably Goodhart's law and the Lucas critique. Here I describe these observations and argue that there is evidence that they apply to IR, thus showing that blind pursuit of performance gains based on optimisation of scores, and analysis based solely on aggregated measurements, can lead to misleading and unreliable outcomes.

1 Introduction

The goal of research on methods for search is to make retrieval systems better at servicing human needs. Variants on this statement of aim might include, for example, 'better at helping humans to complete information tasks', 'better at providing useful information', or 'better at assisting humans by provision of timely and pertinent knowledge'. None of these identifies a quantifiable goal, but arguably they capture the intention of this research even if they are not precise wordings that express it in exact terms. Other researchers might offer different wording but, hopefully, would agree that the above statements are reasonable. That is, it can be contended that there is a broadly understood qualitative aim.

Experiments on search methods are intended to assess the extent to which this aim is being achieved. They can take a range of forms; in particular, batch experiments make use of collections of queries and documents coupled with human relevance judgements. The performance of methods is assessed by using these resources in conjunction with quantitative effectiveness measures, which have a history of development that began over 60 years ago (see van Rijsbergen [1979], Chapter 7 and Saracevic [1995]) and continue to be characterised, refined, and analysed. The kinds of research into how to quantitatively assess search methods include proposals for a wide range of measures, arguments for how specific measures relate to human needs and behaviour, debates over whether measures have desirable mathematical properties, and consideration of how to determine which measures are the most reliable.

In this essay I consider the impact on the gap (or lacuna) between the qualitative aim and the approximation to it that is encapsulated in quantitative measures. This is the basis of what I describe here as the lacuna hypothesis: namely, the well-understood problem that measures are inherently incomplete because the lacuna is fundamental to the challenge of attempting to represent a rich, imprecise human goal in an exact form.

The fact that measures are incomplete as a description of the entity being assessed – in other words, the lacuna hypothesis – is recognised as a foundational consideration in disciplines as diverse as education, climate sciences, bibliometrics, psychology, and economics (see for example Boudinot and Wilson [2020], Flake and Fried [2020], Rowlands [2018], and Strathern [1997]). This incompleteness has led to thoughtful criticisms of the practice of blind pursuit of goals based on measures. In particular, these include Goodhart's law [Chrystal and Mizen, 2001], often stated as 'when a measure becomes a target, it ceases to be a good measure', and the Lucas critique [Lucas, Jr., 1976], which can be condensed as 'predictions of relationships between measurements based on correlations observed in aggregated data are unreliable if the causes of the relationships are unknown'. Both of these arose in economics but have more general applicability.

In their original contexts, Goodhart's law and the Lucas critique formed the basis of arguments that showed that neglecting the lacuna hypothesis can lead to perverse and unintended outcomes in economic policy. Here, I argue that those same effects can be seen in research on mechanisms of search, or retrieval methods (RM), a category that includes similarity metrics, learning-to-rank, and approaches to document indexing. My focus is on RM but the issues I consider are present in other aspects of search, such as interfaces, query expansion, and user experience, and also cognate areas such as summarisation, clustering, word embeddings, filtering, recommendation, and query completion. Improvements in recommendation and filtering, for example, are assessed by use of measures that are of much the same abstract structure as the measures used for search. The same kinds of measure are also used in areas beyond IR such as machine learning, natural language processing, and image recognition.

This essay rests in part on observations in my own past papers over the last two decades, integrated here through the perspective of the lacuna hypothesis. In particular, one of these papers explored how lack of attention to the distinction between qualitative and quantitative aims could lead to meaningless research outcomes, using duplicate detection as a case study [Zobel and Bernstein, 2006]. As in that previous work, this essay is not an argument that specific effectiveness measures are flawed, or that one accepted measure for batch assessment is superior to another. These are serious issues; for example, it is known that some measures can behave in undesirable ways and that there are common poor practices exhibited in experimental designs. However, they are distinct from the problems considered here.

Rather, this essay is a broader argument: the way in which effectiveness measures are used is in general oblivious to deep issues of principle that limit the value of measurement. Specifically, I argue that Goodhart's law, the Lucas critique, and other related observations are directly pertinent to research on RM, in particular with regard to two common practices: of developing search methods purely based on measured scores without reflection on the behaviour that the methods are meant to be supporting; and of basing conclusions on aggregate results across sets of queries without consideration of whether the aggregates are reflective of individual query performance.

Appreciation of the lacuna hypothesis and the observations that rest on it can lead to better experimental practices and give confidence in measured outcomes. Without an understanding of these observations, researchers can unknowingly report that an innovation is an improvement when in fact it is not an improvement at all, an issue that can apply to an individual paper but more concerningly across large-scale programs of work, with the potential for phantom advances and misleading outcomes.

2 Batch measurement of retrieval methods

In the standard approach to batch assessment of RM, sometimes known as offline assessment, there is a collection of documents and queries. For each query, some of the documents have been judged by humans, yielding a relevance assessment; the relevance value of the remainder is unknown. A method is tested by using it to execute each query against the collection, giving a ranking per query of documents, each of which is then translated to a run of relevance values. The aim is to develop methods that place the relevant documents as densely as possible at the start of the ranking, where 'densely' also means that they tend to be ordered by decreasing relevance score.

Effectiveness measures are functions that translates a run into a numerical score [Moffat, 2013; Sanderson, 2010]. All proposed measures have the same broad property of scores being high for runs in which there is good density at the start. How they vary is in their response to defects in the density, that is, the ways in which irrelevant or low-relevance documents are mixed in among the relevant and high-relevance documents. Some measures are very simple, such as binary reciprocal rank, where the score depends only on the position of the first relevant document, or precision-at-k, which considers only the number of relevant documents in the first k positions. Others are much richer; examples include normalised discounted cumulative gain, average precision, and rank-biased precision, but there are many more. Some measures were established by precedent – that is, proposed without a strong rationale but adopted because of the work in which they had been used – and others through academic argument.

There is an extensive literature proposing and analysing these measures; again, see Sanderson [2010], noting that many dozens of papers have appeared since Sanderson's review was completed. From the perspective of this essay, a key component of this literature is the considerations of how well specific measures address the qualitative aim of assisting the user, or how well they simulate 'users of a searching system in an operational setting' [Sanderson, 2010]. Over the last twenty years or so, this has become an explicit goal of the design of measures, exemplified in a paper by Moffat et al. [2017], in which there is a contrast between different ways of defining the user's goal, each of which is characterised by parameterisation of the measure.

What this work illustrates is that the human behaviour being modelled by effectiveness measures is both generalised – all users are characterised in the same way – and highly abstract. It would be astonishing to discover that some researchers regarded these abstractions as a sufficient description of humans; they are obviously incomplete and it is accepted and understood that there is a gap between these computational representations and the extent to which RM are successful at supporting real-world activities.

There is also a body of research exploring whether and to what extent measures correspond with reported human experience or with human assessment of performance; see for example Huffman and Hochster [2007]; Moffat et al. [2013]; Sakai and Zeng [2021]; Turpin and Hersh [2001]; Turpin and Scholer [2006]; Wicaksono and Moffat [2020]; Zhang et al. [2020]. These found evidence of correlation, if weak in some cases, but also showed that the measures remain highly approximate as estimators of human experience.[1] That is, this research demonstrates that the gap between machine and human is present. This is related to but distinct from the question of external validity of results, that is, the extent to which they generalise to new contexts, which was noted as a concern by Bailey et al. [2015] but has had little explicit consideration.

An underlying question is of why measures are used. According to representational theory, measurement concerns 'the correlation of numbers with entities that are not numbers' [Michell, 2021, citing Russell, 1903].[2] It is a truism of science that the purpose of measurement is to confirm hypotheses and thus provide predictivity, that is, to establish that the phenomena under study are likely to have the claimed properties in future observations. In RM, this aspect of measurement has had some attention in the literature on effectiveness measures and significance; for example, see Cormack and Lynam [2006], Rashidi et al. [2021], and Webber [2010]. However, I am not aware of a methods paper in which the predictivity of the results is considered in the analysis.

Implicitly, in most papers measurements are a claim that if one method is superior to another on a test collection (or suite of test collections) under an effectiveness measure (or suite of measures) then the same superiority will be observed in general. Measurement is also used in research in general in a more limited way, to assess the properties of a case study, as opposed to a study in which, as above, the claims are more sweeping. The study-versus-case-study distinction is not widely made in IR, however.[3]

However, another unstated element in most IR papers is the scope of the prediction. That is, it is rare for authors to state the conditions under which their hypothesis will hold. The usual IR test collections have some diversity, with open Web data, newswire, and research articles, but are poor for challenges such as site search, technical material, support of professionals, emails, clinical practice, patent search, organisational collections, longitudinal health records, and essays – in some cases because the test queries are limited, in other cases because the documents are unrepresentative or simply because material of that kind hasn't been curated into a collection.

[1] Which is not to say that the online experiments (user studies) are necessarily a gold standard. They relate to specific study cohorts and tasks and are subject to a range of issues such as representativeness of the participants and constraints on the feedback that participants can provide; if they were perfect it would not be so difficult to do them well [Kelly, 2009]. Thus they are an alternative mode of evaluation, but also have limitations that are similar to those that are well recognised in psychological or clinical studies [Maul, 2013].

[2] The validity of representational theory is contested, as Michell explains, but this high-level statement stands as a succinct common-language description of the core concern. In my view, Michell's introduction to the debates of measurement theory applies as well to RM as it does to his discipline, psychology. See also https://plato.stanford.edu/entries/measurement-science/, accessed 30 March 22.

[3] When I made this claim during a presentation, audience members offered anecdotes concerning the fate of case studies they had undertaken – work suggesting lines of investigation, showing unexpected behaviour, refuting previous claims, and so on – that sounded both reasonable and promising. A consistent theme in these anecdotes was of reviewers who rejected the papers for focusing on a single data set or not undertaking a larger-scale study, apparently not appreciating the value that a case study can offer.

That is, there isn't an agreed way of describing what is being claimed; indeed, what is claimed is rarely explicit. Often, all that is asserted is that a score improvement has been achieved – on specific data sets whose limitations are not examined. Many IR papers could reasonably be described as having experimental designs that are entirely formulaic, without consideration of why the formula is being followed. Such papers can seem to embody an approach in which the data set defines the problem, without identification of a higher goal.

Moreover, although the behaviour and nature of search clearly depends on the search task, the claims for search techniques are implicitly for search in general – an inherent contradiction that has largely gone unexamined. (There are exceptions, however, such as Sirotkin [2012]'s reasoned development of a measure for a specific task, in explicit contrast to the measures in general use.) These issues further highlight that quantitative measurement of effectiveness, as currently practiced, is at best only a broad approximation to the qualitative aims.

At the core of RM batch experiments, then, there is an innate gap between user experience and measurement scores. Its existence has been assessed in comparisons of results from user studies and batch experiments, but as far as I am aware this gap has not previously been explicitly considered as a confound in its own right, and is not noted in overviews of effectiveness measurement including Sanderson [2010], Voorhees [2001], or Manning et al. [2008], Chapter 8. This essay is a reflection on what problems the gap might lead to.

A range of other issues in batch effectiveness measurement have been explored. They are out of scope for this essay but are noted here to contrast them with the topics I do consider:[4]

- Use of multiple correlated effectiveness measures as if they were independent of each other [Webber et al., 2008];
- Unreflective use of particular measures, in particular average precision, despite ongoing debate and a lack of agreement on the behaviour they represent [Moffat and Zobel, 2008; Moffat et al., 2013]; and relatedly,
- Use of measures that are designed for deep pools when the judgements are shallow [Lu et al., 2016, 2017];
- Use of the concept of recall in contexts such as the Web where it is effectively meaningless [Zobel et al., 2009];
- Use of weak baselines, thus inflating the apparent improvement due to the new methods that are being proposed [Armstrong et al., 2009];
- Relying on query sets that may be unrepresentative of the likely query population or uses of the document collection [Hawking et al., 2009];
- Lack of acknowledgement that there is inherent imprecision in effectiveness measurement, due for example in variation in the documents included in the collection or to incompleteness or shallowness of the relevance judgements [Moffat and Zobel, 2008; Moffat et al., 2018; Zobel and Rashidi, 2020];
- Use of inappropriate statistical methods such as Pearson's correlation for top-weighted rankings [Webber et al., 2010];
- Poor use of significance testing [Sakai, 2016];
- Lack of recognition that the predictivity of experiments may be limited, and that some measures are less predictive than others [Rashidi et al., 2021]; and relatedly,

[4] Noting that this is primarily a list of criticisms made by myself and collaborators, not an exhaustive survey.

- Lack of consideration of the scope (or limits on the scope) of the conclusions, in particular with regard to the kind of collection – Web, microblog, research papers, clinical records, and so on – to which they might apply [Rashidi et al., 2021].

Not all of these issues are generally accepted, but they do illustrate the breadth of concerns, and collectively suggest that the accepted methodologies are in some respects naïve. As is clear, though, they are issues with the measures themselves, not with the general principle of the relationship of measurement to a qualitative aim.

3 The lacuna hypothesis

The existence of a gap between computed metrics and real-world goals is recognised as a concern in many disciplines, such as education, psychology, climate science, and economics. A common terminology for a measure of this kind is a *proxy*, also known as an indicator, which is understood to be an inexact substitute for a value, construct, or condition that cannot be observed.

Proxy: A numeric score that is intended to approximate a qualitative aim.

Proxies are used when there is a true value but it cannot be directly measured, as for example is the case for the temperature of molten rock in the interior of the Earth; or when there is no numeric value to estimate, as for example is the case with the competence of a student in a university subject. For the former, a proxy might be a near-surface measurement coupled with inferences from a geological model. For the latter, which is more similar to the use of proxies in RM, the proxy might be the score achieved in an exam.

Another example from education is the capability of a teacher. Student-feedback scores are widely (and controversially) used as an indicator of teaching quality. As an indicator, they have defects, as illustrated by the wide range of scores typically given by individual students as well as the fact that the same teacher can get different scores on different subjects and in different semesters. They can be diagnostic of both excellence and severe failings and thus do have some value; but like any weak indicator there are obvious risks to relying on them as a source of truth. Similar arguments can be made for citations, or citation products such as the H-index, which are used as proxies for academic excellence. Their shortcomings are well-known; they are an indicator, but an unreliable one.

The use of proxies in economics has similarities to that in RM. Gross domestic product is a proxy for standard of living, itself a qualitative aim that is not well understood: 'how much a country produces' is highly indirect as an indicator of 'how comfortable life is'.[5] Another example is inflation, where calculations are based on weighted averages of changes in price of the items in a basket of goods whose contents shifts over time, and where the items in the basket can be varying in price in drastically different ways and between different parts of the country. It is superficially objective but is in fact a manually chosen basket of separate products whose price changes have a variety of causes. If inflation is interpreted as representing the qualitative aim of 'how much more does a household have to spend to maintain the same standard of living', then the answer

[5] Brynjolfsson et al. [2019] explains GDP but additionally uses a consumer study to show how digital goods, which GDP omits, also affect our perceived well-being – thus demonstrating a limitation of GDP as a proxy.

will vary greatly between households and between income brackets; on this interpretation, it can even be the case that no household or item is typical of the average.

For these proxies, abstract values are calculated from mass data across a wide range of different environments but are taken as an indicator of a single overall state. However, proxies are well understood to be only approximations to a qualitative aim, or rather, to the extent to which a qualitative aim has been achieved. Here, I suggest the word *karpos*, variously defined as harvest, yield, utility, or return.[6]

> *Karpos*: The degree of success in achieving a qualitative aim.

The aim, then, is to identify proxies that are good indicators of *karpos*, while acknowledging that there is an inevitable gap. Here, I refer to this gap – again noting that there is good awareness of it across many fields of research, that is, I am not claiming that the concept is original – as the *lacuna*, or absence or void.[7]

> *Lacuna*: The gap, or distinction, between the quantitative and qualitative aims.

The *lacuna hypothesis* (again, my words) is that the gap cannot be closed: for endeavours rooted in the human world, involving human actions, reflecting human behaviours, and so on, proxies are innately inexact. The inevitability of the lacuna cannot be proved – hence it is a hypothesis, not a theory – but intuitively the correctness of the hypothesis is compelling.

This terminology helps to clarify discussion of issues with effectiveness measurement in RM. For example, it isn't obvious (to me) what the *karpos* is that the proxies are meant to correspond to; indeed, it isn't obvious that there even is a single qualitative aim, but rather several interrelated but distinct aims that are not generally distinguished from each other.[8] An attempt to articulate some possibilities was given in the opening paragraph of this essay's introduction.

A potentially surprising aspect of the *karpos*–proxy relationship is that it is asymmetric: improving *karpos* should lead to better measured scores, but improvements in scores, particularly from a strong baseline, can easily not lead to improved *karpos*. An approach to understanding this asymmetry is that the *karpos*-to-proxy direction should hold by construction: if a proxy is of value it should reflect *karpos*. However, in the other direction a proxy can be seen as a parameterised function whose value changes in complex ways (the parameters are those of the retrieval method being measured), with, it is to be expected, many local maxima; only some of these will relate to *karpos* improvements.

In the absence of a qualitative aim, it is not easy to assess whether effectiveness measures have good fidelity, that is, do reasonably well at approximating the *karpos*. Some measures have only low fidelity to the task – much as might be the case for, say, using colour to evaluate the tastiness of chocolate or total vehicle weight to assess the value of a car. Potentially, RM measures are being chosen for their legibility, that is, the ease with which they can be understood or communicated;

[6] While this concept is implicit in the literature on this topic, I am not aware of an accepted term for it.

[7] As is the case for *karpos*, I am not aware of a general term from the literature.

[8] Sirotkin [2012] explains this issue as 'the problem that tends to plague most search engine evaluations, namely, the lack of a clear question to be answered, or a clear definition of what is to be measured. A metric might be a reflection of a user's probable satisfaction; or of the likelihood [they] will use the search engine again; or of [their] ability to find all the documents [they] desired.'

without an understood qualitative aim, even something as simple (and legible) as reciprocal rank can be regarded as being as informative as, say, discounted cumulative gain. The fact that the same measures are used for a wide range of tasks on a wide range of kinds of data suggests that fidelity may indeed be secondary to legibility. Debates over the value of RM effectiveness measures can be regarded as arguments about their fidelity.

Measures are not the only proxies in RM. In particular, relevance is a proxy for 'usefulness' of a document, or perhaps the 'value' of its content – concepts that have very different meanings in different contexts – and reduces a complex entity with diverse content to a single value. Relevance is arguably better understood than are effectiveness measures, though, as in many experiments it is explained as part of the instructions to assessors. Nonetheless, the meaning of relevance changes in hidden ways between applications or tasks, and it is a highly simplified abstraction; as for other proxies it would be a mistake to conflate it with the rich human constructs it relates to.

4 Goodhart's law

In a range of fields of human endeavour, it has been observed that pursuit of high measured scores may not achieve underlying desired goals. This is known as

> *Goodhart's law*: 'When a measure becomes a target, it ceases to be a good measure',

a wording that is due to Strathern [1997] rather than coined by Goodhart.[9]

Goodhart's law can be seen as a consequence of the lacuna hypothesis. It implies that improving the proxy score does not necessarily improve *karpos*, a statement that seems almost too obvious to be worth saying, but if it is indeed obvious then the general neglect of this issue in IR is even more concerning.[10]

There are two broad factors that underlie Goodhart's observation.[11] The first factor is that tuning a method for performance on a proxy (in RM, an effectiveness measure) does not close the lacuna, but rather can tend over time to accentuate it – that is, the aspects of the system that do not address the lacuna will not improve and thus may become more prominent. The second factor is that any measure can be gamed by people with an interest in reporting a high score, and thus the aims that the measure are intended to represent become subverted.

Some examples were given above: teaching scores for lecturers, manipulation (by governments) of the prices that are used to calculate inflation, or citation indices for researchers, an instance that has been explicitly examined as demonstrating Goodhart's law [Fire and Guestrin, 2019].

[9] Other authors have made similar observations to that of Goodhart; for example, Campbell [1976] observes that 'when test scores become the goal of the teaching process, they . . . lose their value as indicators', a statement that is known as Campbell's law. (Whether Campbell's law and Goodhart's law are different is unclear to me, but these two observations appear to have been made independently.) The observations and criticisms made by Campbell, on a wide range of issues from population control and tax incentives to use of breathalyzers and crime-solving rates, remain as pertinent today as when the article first appeared.

[10] In our context, Goodhart's law explains the flaws in the McNamara fallacy, `https://en.wikipedia.org/wiki/McNamara_fallacy`, accessed 9 April 2022. In its simplest form, it is the assumption that only the things that can be quantified are worth considering – an assumption that is often so obviously wrong, and yet seems to be embedded in many practices. If only the observables are of interest, then the qualitative aim is neglected.

[11] Here I'm using 'observation' in the same sense as, and instead of, 'critique' or 'criticism'.

Strathern [1997] examines the use of ratings across higher education, illustrating why the culture of counting and measurement persists when it is recognised to be harmful, for example through not rewarding other aspects of system or behaviour that are also desirable.[12]

A question then is, what might be the manifestations of Goodhart's law in RM? In my view, a strong candidate is a result I was involved in that found clear evidence that there was no improvement in TREC participating systems over a period of 12 years [Armstrong et al., 2009]. Recall that in each year of TREC, participants each evaluate the provided query set on that year's document collection, thus generating the runs on which relevance judgements are based [Voorhees and Harman, 2005]. These runs are therefore blind, in that participants do not have an opportunity to tune their systems to maximise scores on that data, though they do have access to previous corpora to use for system refinement. All participants have a strong incentive – their reputation! – to develop systems that will outperform those from other teams.

By applying a consistent collection of systems to nine years of TREC data, from 1994 to 2005 (with a gap in 2000–2002), the performance of the participating systems as calculated by average precision – the predominant measure at that time – could be benchmarked against a standard. Results are shown in Figure 1.[13] Here, a result of 0.5 would be obtained by a participating system whose score was the average of the benchmarking scores on that year's data; the normalisation adjusted all scores into the range 0.0–1.0. In the figure, the trend in median, third quartile, and maximum score are shown as lines, with the intriguing result that the median stayed close to 0.5; noting the potential confound that the median may reflect the inclusion of faulty systems.

More startlingly, the best result was observed in 1994; an optimist might claim that there is a suggestion of overall improvement in the 2000s, but it is at best small. This result can be taken as clear evidence that 1994–2005 saw no collective improvement in academic RM, despite the period seeing vast technical change in the field – it covers the arrival, growth, and then dominance of search engines in public life.

Arguably, Goodhart's law predicts that the systems, collectively rather than individually, would be unlikely to show improvement on new data – and indeed this is exactly what happened. My contention is that this result is an illustration of what can occur if the lacuna hypothesis is not considered when methods are being developed; noting too that these TREC results represented best efforts not just of a team but of the peak of the IR research community.

By 1994, the third year of TREC, the practice of tuning systems to score well on the previous year's data was already well established. As a researcher who was deeply engaged in development of participating systems at that time, my experience is that there was often a focus on measured score, with for example fine-tuning of parameters to lift that score – that is, with optimisation of proxy, not *karpos*.[14] In this period, I was aware of intriguing innovations that were tried but

[12] A disturbing example is explained in depth by P. Taylor in 'Rigging the death rate', *London Review of Books*, vol. 35, no. 7, 2013, available at https://www.lrb.co.uk/the-paper/v35/n07/paul-taylor/rigging-the-death-rate. In this case, a hospital undermined reporting of mortality statistics by tweaking the labels used to record patient outcomes, thus changing their position in league tables of hospital competency. While there were valid concerns over the applicability of the methods used to compute the statistics, the hospital's decision to distort the outcomes by deliberately abusing the reporting is an excellent case study of Goodhart's law in action – although the law isn't mentioned.

[13] This figure is a simplified presentation of results shown in more detail in Figure 1 of Armstrong et al. [2009].

[14] Acknowledging that this is a broad generalisation, and contrasting behaviour also occurred, and occurs, such as investigation of specific queries that led to poor scores on all systems.

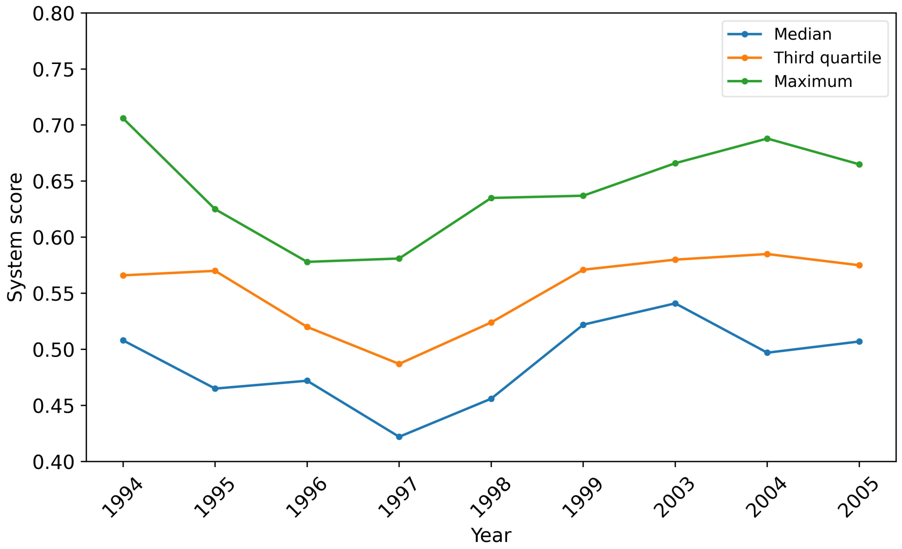

Figure 1: Performance of contributing TREC systems over 1994–2005 (omitting 2000–2002), standardized against a basket of public domain systems as explained by Armstrong et al. [2009]. As can be seen, there is no evidence of improvement over the period.

quickly discarded if their scores were unpromising, even if they were intuitively appealing in terms of the task of search and manual inspection of cases showed that the retrieved documents seemed reasonable. This too is a prediction of Goodhart's law: improvements can be neglected if they do not fit the system of measurement.

The relationship between measures of specific search methods and qualitative goals has been explored, as noted earlier, though from the point of view of this essay leaving questions open. An early pair of papers of this kind [Turpin and Hersh, 2001; Turpin and Scholer, 2006] used a single effectiveness measure to examine whether score-focused degradation of results lists affected user perceptions of results; no real correlation was found, suggesting that scores were uninformative.

In contrast, a recent paper found good correlation between a wide range of measures with assessor judgements of result-page quality [Sakai and Zeng, 2021]. The measures varied in the extent of agreement; some were strong but notably no measure was fully predictive of human judgement. Another recent paper found that measures could be to a limited extent be tuned to improve correlation with reported experience, per-query [Zhang et al., 2020]. Yet another recent paper also found correlation between measures and satisfaction, again finding that some measures were better than others and also, significantly, linking two purposes of measurement: one of human behaviour, the other of human satisfaction [Wicaksono and Moffat, 2020]. Note though that none of these papers investigated whether score differences arising from system changes correlated with improvement in experience – the issue that is the subject of Goodhart's law. However, these papers are evidence for the importance of Goodhart's law in RM, as they found that even with tuning or choice of measure the correlations remain far from perfect.

A question then is, what else might Goodhart's law predict for RM? An obvious candidate is that researchers who are focused on proxy rather than *karpos* might fall into a pattern of following a templated experimental design that examines only improvements in score without consideration of whether the method changes were meaningful in terms of behaviour in practice or value to users, and would regard such score improvements as sufficient in themselves as a result. With time, the body of work as a whole – and the practice of research in the field – could tend to follow this pattern, leading to a collection of papers that report individual improvements that aren't obviously linked to real change in RM from an external perspective. Further, as the score improvements would be detached from perceived value, the correlation between user studies and batch assessment might over time become weak. My contention is that a great deal of published research in RM does indeed have this character.

5 The Lucas critique

Goodhart's law concerns measurements that are intended to approximate a qualitative aim. Another observation from economics is the *Lucas critique* [Lucas, Jr., 1976], which concerns the challenge of determining how measurements of different entities are related to each other.

The Lucas critique does not have a standard description stated in general terms, in contrast to Goodhart's law. As a possibility, I suggest the brief statement

> The *Lucas critique*: Predictions of relationships between measurements based on correlations observed in aggregated data are unreliable if the causes of the relationships are unknown.[15]

The critique can be understood as follows. It concerns a system that has properties (or entities, or observables) that can be measured, and a set of inputs, some of which can be externally controlled. Over time, the measurements can be collected and correlations amongst the entities can then be identified. However, it does not follow that directly manipulating one of the entities will cause the change in the other entities that the correlation would predict, because it is not known what other hidden elements will be affected by the manipulation or whether the correlation was due to a hidden cause.

For example, it might be observed that the amount of electricity used for air-conditioning is correlated with ice-cream consumption. However, this does not mean that banning air-conditioning will reduce the consumption of ice-cream, because both are due to a separate cause – hot weather. Further, the presence of humans in the system means that they may respond by making unexpected changes in behaviour, such as increased use of petrol to cool themselves in cars. That is, it can't be assumed that other parts of the system will remain static. Uninformed interference in the system has unpredictable effects.

Here, total electricity usage and ice-cream consumption are highly aggregated values, reducing the behaviour of whole populations to a couple of variables. Inflation was noted earlier as a highly aggregated index; the standard illustration of the Lucas critique is the correlation between high inflation and high employment. Due to this correlation, there are instances of governments deliberately increasing inflation in order to improve employment, attempts that failed because of the unanticipated human behaviours their actions provoked [Lucas, Jr., 1976]. As another

[15] https://en.wikipedia.org/wiki/Lucas_critique, accessed 4 April 2022, gives 'it is naïve to try to predict the effects of a change in economic policy entirely on the basis of relationships observed in historical data, especially highly aggregated historical data'.

example, a large increase in the cost of fuel – with the intention of reducing fuel consumption – can have the effect of making older, inefficient cars cheaper because they become less desirable. This makes them affordable by people on a low income, and, depending on the mix of incomes in the population, the net effect on fuel consumption might be negligible.

Lucas's observation has two separate implications for RM. The first is that changes in aggregate measured performance that are induced by changes to parameters[16] cannot be assumed to occur again if the relationship between parameter and performance is unknown. In RM, systems are typically developed by observation of performance on aggregate data. This neglects the factors that lead to the aggregate performance, an issue that has been noted in RM (see for example Hull [1996]) but is not often considered. I discuss this aspect of the Lucas critique more fully below.

The second implication is that changes to parameters can change the behaviour of the system, and in particular can change the actions of humans (who, here, are part of the system) as they optimise their behaviour to maximise their personal benefit.

Such issues are to some extent beyond the scope of this essay, as they don't necessarily involve measurement, but they are significant for IR. For example, a change to a search interface to help with query refinement can affect whether the results satisfy the user even if the same information is presented; and can do so even if the suggested query refinements are unhelpful.

More insidiously, consider query auto-completion. It is intended to assist users by reducing the typing effort, but it can instead lead to users entering different queries; and in doing so may undermine one of the uses of query logs, spelling correction, by reducing the variations of queries that are entered. Gaming of auto-completion can lead to drastic changes in the information that users receive – it has been used to trick search engines into offering completions that are racist, sexist, or hateful [Olteanu et al., 2020],[17] and is arguably a variation of the issues predicted by the Lucas critique.

Researchers too are part of a system, of method development, and the actions they take can be influenced by how performance is being measured. As Lucas observed, when their environment changes, people (including researchers) adjust their behaviour to get the best outcome, and as a result the measure and the goal become detached from each other. In this respect, Goodhart's law can be regarded as an interpretation of the Lucas critique.

Returning to the first implication – that discoveries about aggregate behaviour can be ill-founded if the causes of the behaviour are not considered – another observation is that behaviour observed in aggregate might not apply to individuals. This is a form of the well-known ecological fallacy: the assumption that conclusions drawn from statistics about a population apply to all the individuals in the population.

It is a serious error to make this assumption. A simple example is to assume that if on average a medication assists recovery, then it will assist recovery for everyone. More complex examples arise when the data being examined contains a mix of distinct populations, in unknown proportions, or of unknown characteristics. If a parameter is changed, different subpopulations may respond unpredictably; this is another reason that attempts to directly alter the whole can

[16] Using the term broadly; choice of stemming function or of architecture of a neural net are parameters, for example.

[17] See also an excellent overview of problematic autocompletion by C. Cadwalladr published in *The Observer*, 'Google, democracy and the truth about internet search', available at https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook.

fail in unexpected ways. Without understanding or prior stratification of these subpopulations (noting that post-hoc stratification is another experimental fallacy), or of how the parameters affect individual queries, it is easy for modelling to be wrong.

The ecological fallacy is evident in research by Billerbeck and Zobel [2004] in the early 2000s, on automatic query expansion (also known as pseudo-relevance feedback), a technique that is used to add terms to queries to compensate for vocabulary mismatch. In a standard approach to automatic query expansion, first a query would be run in the usual way. The f most significant terms would be chosen from the top k documents returned and the query would then be run again with those terms added. The parameters f and k would be chosen to give the greatest improvement in measured performance, with typical values found to be say $k = 15$ and $f = 13$ – indeed, these were the best values we observed in terms of average score, after allowing both f and k to vary from 0 to 100.

But consider Figure 2, a more detailed variant of the right part of Figure 3 in the original paper.[18] Each different colour represents a single query. Large squares represent the parameter combination where the highest improvement was observed for that query, reducing in size as improvement fell; white (or empty) meant that at that locus the query performance was unchanged or was degraded.

What this picture makes clear is that, across 50 queries, there is no pattern. Some queries improve only for certain k; these are the vertical streaks. The horizontal streaks are those that improve only for certain f. The location $\langle 15, 13\rangle$ is not in any way special, and neither k nor f seems to have much correlation with improvement. In other words, the overall conclusion of 'good' parameter values seems to be entirely without foundation, and would be falsified if the population mix of the queries was changed.

Thus prior conclusions about the best parameter values for automatic query expansion, based on aggregate behaviour, seem to be based on the ecological fallacy; this instance is a strong illustration that aggregate behaviour is simply misleading.

The failure of inference demonstrated in Figure 2 suggests questions that researchers should ask. Are there subpopulations of queries that seem to vary together? Are the results influenced by the chosen effectiveness measure? Is there a relationship to other artefacts, such as the number of known relevant documents? Are poor scores due to unjudged documents – which are common in query expansion because the additional query terms tend to fetch new materials? We observed, for example, that big improvements were often due to the expansion process finding a specific additional term, but that the number of documents and terms needed to ensure that the helpful term was added was unpredictable.

Relatively few papers use per-query examination of behaviour to interrogate the detail of reported results, but those that do often show wide variation with little pattern; that is, the per-query scores are not neatly drawn from a normal distribution. Several examples are shown in recent work by Zobel and Rashidi [2020], which re-examined behaviour for runs participating in TRECs from the mid-1990s to 2004. This work is a clear illustration that the ecological fallacy applies to RM. On the one hand, system rankings and average system scores for average precision were as stable, or more so, than for other effectiveness measures. On the other hand, per-query scores for average precision showed high uncertainty and high variability, while some other measures

[18] This version can be seen in Figure 4.4, page 78, of Billerbeck [2005]; used here with permission.

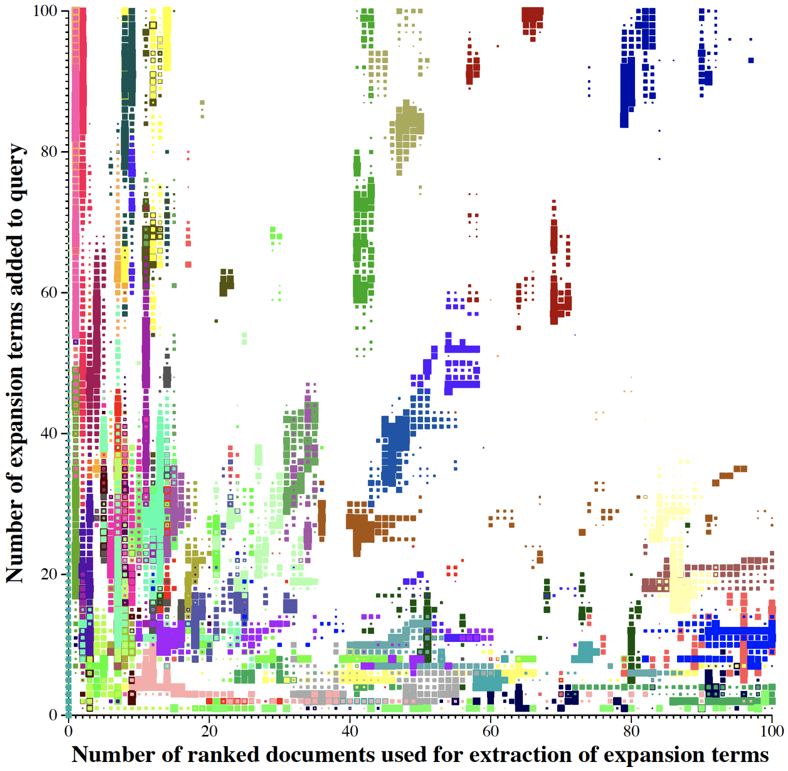

Figure 2: Per-query, the parameter combinations for which each query shows best performance in automatic query expansion; image from Billerbeck [2005], used with permission. Each colour represents a different query (with some colours obscured behind others, but not to an extent that is misleading). As can be seen, the individual queries have highly disparate behaviour.

were much more consistent. There was essentially no relationship between overall and individual characteristics.

Lucas's critique and the ecological paradox demonstrate the need to understand and model individual behaviour, especially outliers that have their own trends. They also highlight the need to recognise subpopulations and issues such as the likelihood of occurrence of particular kinds of query. Results based solely on aggregation can easily be misleading or outright wrong.

6 The goals of research into retrieval methods

Earlier I argued that research into RM lacks a clear or agreed qualitative aim. This issue was the topic of previous essays of mine, as I now review.

Two of these appeared in SIGIR forum. Zobel et al. [2009] presents the argument that there isn't an agreed understanding of recall – a foundational element in many effectiveness measures – and that it is in general meaningless. Zobel [2018] concerns the absence of an agreed understanding of the scope or goals of the field. The lack of agreement is a factor in the issues of lacunae that are the topic of this essay.

This is a long-standing concern in IR; Saracevic [1995] wrote that we need to have 'criteria representing the objectives of the system' and concluded, in effect, that we don't know what they are.[19] A key argument in his paper concerns the need to understand and state assumptions – an issue that is directly aligned with the need to articulate the qualitative aim and in light of that articulation defend the choice of proxy.

The need to explicitly understand the *karpos*-proxy distinction was, in different terminology, the topic of another essay of mine [Zobel and Bernstein, 2006]. Although it was written over 15 years ago, it remains as relevant today: there is still little articulation of qualitative aims and the choice of proxies is almost never defended. Rather, in most papers on RM a proxy or collection of proxies is chosen and reported on without justification or consideration of why they are appropriate to the method or task.

A core of the argument in Zobel and Bernstein [2006] was the observation that in another field of IR, duplicate detection, not only did the same issues apply but in some cases researchers were using the chosen proxy as the goal of the research – even though some of the proxies were simplistic to the point of absurdity. As a slightly simplified example (but only slightly), if one method reported that 10,000 pairs of documents were 'duplicates' of each other, based on a comparison function, and another reported that there were 20,000 such pairs, based on another function, then the second method could be deemed more successful, without any consideration of whether a human would agree that the pairs were indeed duplicates. The reasoning in these papers seemed circular, with the problem being characterised and explained by the outcome.

Another element considered in Zobel and Bernstein [2006] was that in many of the papers on duplicate detection there was no explicit statement, or even implicit assumption, concerning what might define a duplicate. As a concept, it is surprisingly difficult to capture: being identical is too strong a test while mere similarity is too weak; but it is not clear what, definitionally, lies between. Statements such as 'effectively the same document' unhelpfully do no more than beg the question of the meaning of 'effectively'. A small survey of colleagues at the time found near-complete disagreement on the issue of what constituted a duplicate, thus illustrating that the qualitative aim wasn't shared, and certainly isn't simple to articulate. This poses the question of what a similar survey of RM researchers would reveal about the aims of search technologies, and the extent to which they agree – an exercise that, as far as I know, has not been undertaken.

Our work was couched as an application of concepts from the philosophy of science to IR, in particular the use of warrants (formal arguments) to defend qualitative aims and quantitative measures and the importance of paying attention to issues such as fidelity. There is an extensive literature on measurement theory that highlights the ways in which thoughtful people can disagree about the purposes and implications of measurement. However, as insightful as this literature is, it is irrelevant to research that proceeds without even minimal consideration of whether the measures being used are meaningful for the intended task. That is, the subtleties of measurement theory are unlikely to be helpful if basic principles are not being observed.

[19] In this article, Saracevic's focus is on setting out a wide range of questions that should be considered in design of IR experiments, including issues with measures; for example, he highlights concerns with the concepts of recall and relevance.

7 Conclusion

To demonstrate scientific advances, or to make scientific claims, we need to make measurements. In research on retrieval methods, these measurements are proxies for qualitative aims. In principle, it is understood that the goal is to achieve that aim, and not just seek to improve the measured score or to make claims based on measurements that are highly aggregated.

In this essay I have argued that in much of this research there is no demonstrated understanding of the distinction between proxy and aim. Two well-known observations made in the 1970s in the field of economics explain the importance of this distinction, as I have reviewed here. At a high level, Goodhart's law concerns an externality: how measurements relate to reality and why the distinction between measurement and reality is critical to achieving qualitative aims. The Lucas critique, in contrast, essentially concerns internalities: that an incomplete understanding of how systems work can lead to false inferences and incorrect predictions.

The issues that these observations predict are in evidence in IR. For Web retrieval, where the majority of retrieval tasks are simple, and performance modelling is supported by extraordinarily voluminous data, perhaps these concerns are of low importance. However, in other contexts, IR measures are less predictive or robust; there is what appears to be extensive methodological naïvety, with poor practices perpetuated despite significant criticism of them; and potentially there is low agreement on qualitative aims. These issues are widespread in the literature: claims of improvements of systems based solely on measured scores and blind use of aggregation.

As a summary statement, I suggest that researchers should

- *seek to improve the behaviour of a method, not its score* and
- *verify claimed aggregate outcomes by analysis of individual cases.*

This advice has corollaries that researchers – and reviewers! – should be alert to, including:

- Achievement of an improved score does not mean that the method is improved.
- Choice of effectiveness measures that match the aim of the research should be part of the design of the experiment, not an afterthought.
- Results that are based on optimisation to a particular measure should be verified independently of the measure.
- Understanding of accuracy, distribution, and individual variation are critical to accurate interpretation of experimental results.
- A collection of individual cases should not be treated as a homogenous, consistent, or uniformly distributed whole, unless they have been shown to have these properties.

In this essay I have focused on retrieval methods within the field of IR, but similar cultures are evident in other fields, so it is likely that there are similar consequences. That is, this paper can be read as a case study of a problem, but the problem may well be widespread. The use of standardized measures in a simple template is a feature in, for example, natural language processing, machine translation, and machine learning, and thus for these areas as well as IR there are questions to address on the robustness of their research outcomes.

IR has a deep history of exploring measures, with ongoing debates and innovations. We need to continue to ensure that these measures are sound, but it is more important that we use them in more considered ways, with acknowledgement of their flaws and fundamental limitations.

Acknowledgements

I have discussed aspects of this material with a range of colleagues over many years, and thank them all for the fruitful conversations that helped me sharpen my thinking. Most recently, I've valued discussions with Pauline Lin, Alistair Moffat, Lida Rashidi, and Felisa Vázquez-Abad. I also thank Yaniv Bernstein, with whom I first discussed some of the issues presented here.

References

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Manangement*, 2009. doi: 10.1145/1645953.1646031.

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Manangement*, 2015. doi: 10.1145/2766462.2767728.

B. Billerbeck. *Efficient Query Expansion*. PhD thesis, RMIT University, Victoria, Australia, 2005. URL https://researchrepository.rmit.edu.au/esploro/outputs/doctoral/Efficient-query-expansion/9921861213101341.

B. Billerbeck and J. Zobel. Questioning query expansion: An examination of behaviour and parameters. In *Proc. Aust. Document Computing Conf.*, 2004. doi: 10.5555/1012294.1012302.

F. G. Boudinot and J. Wilson. Does a proxy measure up? A framework to assess and convey proxy reliability. *Climate Past*, 16, 2020. doi: 0.5194/cp-16-1807-2020.

E. Brynjolfsson, A. Collis, and F. Egger. Using massive online choice experiments to measure changes in well-being. *Proc. National Academy of Sciences*, 116(15), 2019. doi: 10.1073/pnas.1815663116.

D. T. Campbell. Assessing the impact of planned social change. Technical Report 8, Occasional Paper Series, Public Affairs Center, Dartmouth College, 1976. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.6988&rep=rep1&type=pdf.

K. A. Chrystal and P. D. Mizen. Goodhart's law: Its origins, meaning and implications for monetary policy. In *Festschrift in honour of Charles Goodhart*, Bank of England, 2001. doi: 10.4337/9781781950777.00022.

G. V. Cormack and T. R. Lynam. Statistical precision of information retrieval evaluation. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2006. doi: 10.1145/1148170.1148262.

M. Fire and C. Guestrin. Over-optimization of academic publishing metrics: observing Goodhart's law in action. *GigaScience*, 8, 2019. doi: 10.1093/gigascience/giz053.

J. K. Flake and E. I. Fried. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 2020. doi: 10.1177/2515245920952393.

D. Hawking, T. Rowlands, and P. Thomas. C-TEST: Supporting novelty and diversity in testfiles for search tuning. In *Proc SIGIR Workshop: Redundancy, Diversity, and Interdependent Document Relevance*, 2009.

S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2007. doi: 10.1145/1277741.1277839.

D. Hull. Stemming algorithms: a case study for detailed evaluation. *J. American Society for Information Science*, 47, 1996. doi: 10.5555/231880.231890.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2), 2009. doi: 10.1561/1500000012.

X. Lu, A. Moffat, and J. S. Culpepper. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal*, 19(4), 2016. doi: 10.1007/s10791-016-9282-6.

X. Lu, A. Moffat, and J. S. Culpepper. Can deep effectiveness metrics be evaluated using shallow judgment pools? In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2017. doi: 10.1145/3077136.3080793.

R. E. Lucas, Jr. Econometric policy analysis: A critique. In *Carnegie-Rochester Conference Series on Public Policy*, volume 1, 1976. URL https://EconPapers.repec.org/RePEc:eee:crcspp:v:1:y:1976:i::p:19-46.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

A. Maul. Method effects and the meaning of measurement. *Frontiers in Psychology*, 14(169), 2013. doi: 10.3389/fpsyg.2013.00169.

J. Michell. Representational measurement theory: Is its number up? *Theory & Psychology*, 31 (3), 2021. doi: 10.1177/0959354320930817.

A. Moffat. Seven numeric properties of effectiveness metrics. In *Proc. AIRS Asian Information Retrieval Symposium*, 2013. doi: 10.1007/978-3-642-45068-6_1.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Database Systems*, 27(1), 2008. doi: 10.1145/1416950.1416952.

A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Manangement*, 2013. doi: 10.1145/2505515.2507665.

A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Transactions on Information Systems*, 35 (3), 2017. doi: 10.1145/3052768.

A. Moffat, F. Scholer, and Z. Yang. Estimating measurement uncertainty for information retrieval effectiveness metrics. *ACM Journal of Data and Information Quality*, 10(3), 2018. doi: 10.1145/3239572.

A. Olteanu, F. Diaz, and G. Kazai. When are search completion suggestions problematic? In *Proc. Human-Computer Interaction*, volume 4, CSCW2, 2020. doi: 10.1145/3415242.

L. Rashidi, J. Zobel, and A. Moffat. Evaluating the predictivity of IR experiments. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2021. doi: 10.1145/3404835.3463040.

I. Rowlands. What are we measuring? Refocusing on some fundamentals in the age of desktop bibliometrics. *FEMS Microbiology Letters*, (365), 2018. doi: 10.1093/femsle/fny059.

T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2016. doi: 10.1145/2911451.2911492.

T. Sakai and Z. Zeng. Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures. *ACM Transactions on Information Systems*, 39(2), 2021. doi: 10.1145/3431813.

M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4), 2010. doi: 10.1561/1500000009.

T. Saracevic. Evaluation of evaluation in information retrieval. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 1995. doi: 10.1145/215206.215351.

P. Sirotkin. *On Search Engine Evaluation Metrics*. PhD thesis, University of Düsseldorf, 2012. URL https://arxiv.org/pdf/1302.2318.pdf.

M. Strathern. 'Improving ratings': Audit in the British university system. *European Review*, 5, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3⟨305::AID-EURO184⟩3.0.CO;2-4.

A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2001. doi: 10.1145/383952.383992.

A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2006. doi: 10.1145/1148170.1148176.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.

E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proc. CLEF Cross-Language Evaluation Forum*, LNCS 2406, 2001. doi: 10.1007/3-540-45691-0_34.

E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005.

W. Webber. *Measurement in Information Retrieval Evaluation.* PhD thesis, The University of Melbourne, Victoria, Australia, 2010. URL http://hdl.handle.net/11343/35779.

W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2008. NO DOI.

W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), 2010. doi: 10.1145/1852102.1852106.

A. Wicaksono and A. Moffat. Metrics, user models, and satisfaction. In *Proc. WSDM Int. Conf. on Web Search and Data Mining*, 2020. doi: 10.1145/3336191.3371799.

F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval*, 2020. doi: 10.1145/3397271.3401162.

J. Zobel. What we talk about when we talk about Information Retrieval. *SIGIR Forum*, 2018. doi: 10.1145/3190580.3190584.

J. Zobel and Y. Bernstein. The case of the duplicate documents: Measurement, search, and science. In *Proc. APWeb Australia-Pacific Conference on the Web*, 2006. doi: 10.1007/11610113_4.

J. Zobel and L. Rashidi. Corpus bootstrapping for assessment of the properties of effectiveness measures. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Manangement*, 2020. doi: 10.1145/3340531.3411998.

J. Zobel, A. Moffat, and L. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? *SIGIR Forum*, 2009. doi: 10.1.1.415.6729.