What Matters in a Measure?

A Perspective from Large-Scale Search Evaluation

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

Gabriella Kazai*
Amazon
London, United Kingdom
gkazai@amazon.co.uk

Nick Craswell
Microsoft
Seattle, United States
nickcr@microsoft.com

Seth Spielman
Microsoft
Boulder, United States
sethspielman@microsoft.com

ABSTRACT

Information retrieval (IR) has a large literature on evaluation, dating back decades and forming a central part of the research culture. The largest proportion of this literature discusses techniques to turn a sequence of relevance labels into a single number, reflecting the system's performance: precision or cumulative gain, for example, or dozens of alternatives. Those techniques—metrics—are themselves evaluated, commonly by reference to sensitivity and validity.

In our experience measuring search in industrial settings, a measurement regime needs many other qualities to be practical. For example, we must also consider how much a metric costs; how robust it is to the happenstance of sampling; whether it is debuggable; and what activities are incentivised when a metric is taken as a goal.

In this perspective paper we discuss what makes a search metric successful in large-scale settings, including factors which are not often canvassed in IR research but which are important in "real-world" use. We illustrate this with examples, including from industrial settings, and offer suggestions for metrics as part of a working system.

CCS CONCEPTS

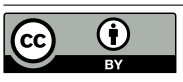
• **Information systems** → **Relevance assessment**; **Test collections**; *Crowdsourcing*.

KEYWORDS

offline metrics; validity; efficiency; sensitivity; reliability

ACM Reference Format:
Paul Thomas, Gabriella Kazai, Nick Craswell, and Seth Spielman. 2024. What Matters in a Measure?: A Perspective from Large-Scale Search Evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3626772.3657845

*This work was carried out while at Microsoft.



SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657845

1 INTRODUCTION

Information retrieval (IR) research has a strong tradition of shared metrics and careful measurement to distinguish a good from a bad system or to measure improvement over time [79, 89]. More recently, such metrics have also been optimisation targets for machine-learned rankers and other IR components.

In either use, differentiating a better from a worse system means committing to some mechanism for measurement. It also means engaging with broad questions about the quality and relevance of information. Choosing a metric locks in one's definition of a good system and in so doing embeds certain costs, risks, and biases. We argue that choosing a definition of quality can have repercussions not only for systems, but also for the organisations that build and maintain them.

There are ways a metric itself can succeed or fail that are not obvious. In this paper, we discuss the design and selection of metrics from our experience in large-scale industrial settings, with an emphasis on Cranfield-style or "offline" measurement of effectiveness. We touch upon statistical properties that a metric might have, but give greater emphasis to other properties not often discussed in the IR literature.

A note on context. The observations and anecdotes below are coloured by our backgrounds. We each have an academic background, mostly (but not entirely) in information retrieval, and in our academic work, we make heavy use of conventional offline metrics and tools such as TREC collections (indeed, we have managed several TREC tracks and data releases). Our background is also in large industrial settings, working in teams that develop and maintain core web search metrics; help design and interpret metrics for other more special-case web measurement problems; and advise on search problems in enterprise settings, again with a view of measurement across the organisation. The notes that follow may not capture what is most useful in other settings: however, we believe most of these considerations need to be addressed in all large-scale search applications.

1.1 Why measure information retrieval?

There are at least three reasons we might be interested in a reproducible, well-defined measure of search quality.

Demonstrating value. The most obvious is to demonstrate value: that is, to show that someone would value what we have built. One measure here is income—if someone is paying for search, they likely value it—but this is a loose proxy at best. Usage is somewhat more connected with value, but still many steps removed from the choices

made in building an information system. A measure of search quality can more directly demonstrate the value, especially the value of an incremental change to a system and/or an innovative research idea.

The simplest way to demonstrate value is with a point measure, giving some sort of indication of how valuable a particular system might be in a particular circumstance. Such a measure could say "if we bought this system, staff could find what they need 80% of the time"; or "this algorithm ranks 1B web pages in 0.1 s". As well as these point measures, we might be tracking a system over time: for example we might care about how search engines cover evolving topics, or whether the system we're building is improving as time goes on, or whether searcher behaviours are changing as devices change. *Validity* and *stability* are important in this case—we need to measure something real, and any variation should be due to changes in the rankings not due to measurement noise.

Point measures such as these are also used to gather data for machine learning algorithms, either for training or to evaluate a learned model. *Volume* is important here, as is adequate *representation* of many classes and corner cases.

Motivating improvement. Once we can talk about value, a second reason to measure is to motivate improvement. By measuring a concrete notion of value, we define what is important; by rewarding improvements on a concrete metric, we drive improvement on those things the metric measures. For example, TREC has historically rewarded systems that return on-topic documents, but not systems that account for reading level or credibility, the result in that forum has been improved models of topicality, but little work on credibility. This suggests more than one metric may be needed for most systems, as (for example) we can't use a measure of topicality to motivate improvements from a UI design team nor use a measure of interface efficiency to reward improvements from an image indexing project.

An important instance of improvement comes after some experiment (or quasi-experiment) has run its course, and we need to make a decision. We might need to decide whether to purchase system A or B, or whether to update our search engine with some new algorithm, or whether we need to intervene in some problematic case. In these situations, the *sensitivity* of a metric, its ability to detect changes of a fixed magnitude, is important as well as the criteria above.

Communicating. Fundamental to both value and improvement—indeed prior to both—a third reason to measure is that measurement gives us a tool for communication. Concepts like "better search" are vague to the point of being useless; a well-defined measurement forces us to agree on what "better" is, how it is manifested, what is being traded off, and how we know "better" when we see it. Having a reliable, agreed, metric lets us debug, develop, and plan with a common language. For this, a metric needs to be *interpretable*.

None of these uses and none of these desiderata are unique to large-scale web search or other commercial applications: we would argue that small-scale experiments, and academic research, use metrics for the same reasons. In any case, it is worth remembering that *the metric is not the goal*. We do not measure precision because we care about precision; we care instead about some business goal, or at least we should. This larger goal might be to reduce cost or time, for example; to grow market share or revenue; or, more abstractly, to produce more relevant or more useful results.

An edible analogy. As an analogy, consider a restaurant. A customer might indicate they'd like a dessert: the restaurant's job is to satisfy this partly-articulated need. The restaurant may offer a menu, representing what's available and dessert-like; the customer can then choose to commit to one of those.

As the food comes out of the kitchen, the chef will be at the "pass", looking at the dish to ensure quality. The chef might check, first, that it is as described or as ordered—is the chocolate ice-cream actually chocolatey? The chef might also choose to check that the restaurant's other polices are met, even if they're not specified in the menu: the food is at the right temperature, the plate is the right size, etc. A careful chef may check things a customer would never notice: kitchen cleanliness for example, or where ingredients come from. A careful and somewhat nosy chef could also gather data after the dish has been served, by seeing whether the whole thing was eaten or whether it was sent back.

Of course using the right plate, or getting the order correct, is important and may even be how kitchen staff are judged and paid. It is not, however, the restaurant's goal. The restaurant's goal is to have a paying customer; hopefully a happy customer; ideally even a repeat customer, or one who'll recommend the restaurant to their friends. There are some aspects of a dish which are easy to check, and we hope that attending to them leads to happy diners; and we can measure revisits, and trust they have something to do with the aspects we measure; but they are not the same thing.

The parallel is clear. The request for a dessert is a query, the menu a SERP, and committing to a dish is selecting a result from the SERP—hopefully one which was represented faithfully in the first place. Our metrics are like the chef at the pass: we might check topicality or relevance (what was ordered), and we should also check other policies (quality, timeliness, curatorial decisions). We could also look at searchers' behaviours once they have a result. None of these checks are the goal in themselves. Getting the topic right, or serving timely information, should correlate with a satisfied searcher; but there will always be more to the story. We must not confuse the metric with the goal.

1.2 How to measure

How then can we measure search effectiveness? Kelly [48] gives a range of methods, from the "system focus" of TREC ad-hoc to the "human focus" of studying behaviour in context, via log analysis and interactive studies amongst other methods. Zangerle and Bauer [97] similarly divide between "offline", "user study", and "online" methods. Zhai meanwhile divides between methods assessing the absolute utility of a system, via A/B tests or small-scale lab interactive studies, and assessing relative performance via TREC-style work [98].

This one-dimensional view of metrics hides a number of interesting design choices. Our metrics rely on some sort of performance signal, and so we offer another classification based on three characteristics of these signals (Table 1):

(1) *What* signal is offered? Is it implicit—e.g., derived from behaviours such as clicks—or an explicit statement about a system's performance, such as from a survey or complaint?
(2) *Who* is providing the signal: the searcher themselves, or some third party? (These are the "gold" and "silver" or "bronze" assessors of Bailey et al. [7].)

Signal source		Signal timing: During use	Signal timing: Post-hoc
Implicit	Searcher	"online", e.g. clicks	e.g. use of result
	3rd party	—	e.g. time to judge
Explicit	Searcher	e.g. feedback tools	e.g. questionnaire
	3rd party	e.g. think-aloud	"offline", "Cranfield"

Table 1: Examples of techniques for gathering signals of search performance. In this work we concentrate on explicit labels from third-party judges in the offline or "Cranfield" model, famously described by Cleverdon et al. [18] and exemplified by TREC [89].

(3) *When* is the the signal available: at the time the search is carried out (during use), or some time later (post-hoc)?

These signals are mediated by the interface and instrumentation.

There are several possible combinations of characteristics. Most of these combinations have been explored, in research or in practice.

Implicit feedback from the searcher, post-hoc, could include observing how a search result is used; this can be useful for example in private settings or where we have instrumented tools [51, 86]. Explicit feedback from the searcher could include the use of feedback tools on a SERP (during use) or questionnaires after the event (post-hoc) [52]. Implicit signals from third parties include the time taken to form a judgement [95]. Less well-used due to the relative expense, but a source of rich data, are explicit judgements from third parties during use: think-aloud lab studies are a useful method of this kind [67].

Each of these combinations can give rich data, but are relatively expensive and/or sparse, and are hard to reuse. The remaining two combinations, on the other hand, are very commonly used. "Online" methods use implicit feedback, gathered from searches in situ [49, 54, 60, 97]. This feedback could include signals such as click position and rate, dwell time, scrolling, or query reformulation. There is some evidence that these online signals are valid, in that they seem to correlate with expressions of emotion and other feedback [23, 61, 62].

Explicit feedback, collected post-hoc from third parties (i.e. not from the searcher), is typical of the "Cranfield" or "offline" approach[1]. Their relatively low cost and support for fast experimentation means offline metrics are heavily used in both industrial and academic practice, indeed with very similar techniques in our experience. We will focus on offline metrics in the remainder of this paper.

1.3 Offline metrics

Understood this way, measuring with an offline metric requires at least the following components:

Objects to evaluate. Typically, these are several individual results (documents) for each of several queries or information needs, sampled from a representative set. Getting these objects involves methods for sampling representative (or otherwise interesting) queries, and generating results for these queries in a repeatable way that replicates what a searcher would experience. This latter requirement constrains how we sample: for example, sampling web search results that include advertisements predicts success better than focusing on "organic" results alone [84].

A labelling scheme. Some process is needed such that each object can be labelled for its quality. Note that this includes a set of value judgements, i.e. some preconceived notion of what is good: this might be topicality, authority, recency, reading level, or some other desideratum. Defining a labelling scheme also means choosing a scale (ordinal, for example, or ratio); and a method for assigning labels with that scale (how many options or points are offered; whether they are all named; whether we collect absolute or relative labels; and so on) [72].

A process for assigning labels to objects. This involves gathering data for the sampled queries, documents, and/or query-document pairs from assessors (in-house, contracted, or crowd). In practical terms this includes building an instrument for assessors that guides labelling; as such, it embodies the labelling scheme. At the the same time, it is a tool used by people and careful design is needed to minimise misunderstanding and bias. Instruments should ideally include instruction or guidelines, and (where crowd assessors are used) should include crowd quality measures as an integral part [2, 70].

A process for metric computation. This is an aggregation: a mechanism for converting a series of labels into a summary number. This could include transformations over labels (e.g., mapping labels to gains), summarising across labels or across objects, or combining labels from multiple aspects [32, 50]. The numeric properties of metrics have been discussed at length [63, 64], as have methods for aggregating multiple labels from independent assessors [22, 24, 90].

A statistical framework. Finally we need some means to decide when a metric represents an improvement, or has changed in some interesting way. In most literature this is a simple t-test over per-query scores, but alternatives are certainly possible [31, 75].

2 WHAT MATTERS IN A MEASURE?

There are, it sometimes seems, at least as many IR metrics as there are IR researchers—there are a great many ways to embody the ideas above. This means that as IR researchers and practitioners we need to decide amongst metrics, and to do this we need to decide what makes a "good" metric [25, 35].[2]

This has been discussed in the literature, where common qualities include sensitivity; correlation other established metrics; robustness to missing data (hence also budget constraints); or correlation with observed behaviours or self-reports. These qualities are still important in an industrial setting, but measuring at scale, repeatedly, and in a large organisation adds other practical requirements. In Table 2 we summarise some of the main concerns from this perspective:

- *Social and legal* aspects, not often considered in academic work on IR measurement;
- *Validity* or fidelity to what we want to measure;
- *Efficiency*, particularly important when working at large scale or high frequency;

[1] "Cranfield" evaluations could use feedback from the searcher, and several well-known collections do in fact do this. In practice, distributed and especially crowd judging has made this less common than third-party judgements.

[2] We need also consider the context of use for the metric: that is, what to do with the metric once it's computed, how and to whom to report it, with what timing, and so on. These issues are important, but are also highly dependent on the setting and are not canvassed here.

- *Reliability and sensitivity*, whether we can read any useful signal at all;
- *Interpretability*, also little considered in the literature but of key importance when using a metric to design or debug a system;
- *Organisational* aspects, distinctive to settings where multiple people or teams are working to improve an agreed metric.

In the sections that follow we expand on each of these dimensions, emphasising what we have found important in our work as applied IR researchers and practitioners.

3 SOCIAL AND LEGAL

We now turn to what makes a measure most useful, starting with the social context. It is well understood that search and similar tools can have substantial social impact [10, 11, 29, 39, 55, 68, 96], but the impact of metrics and metrics design is not commonly discussed. Besides legal, regulatory, or policy constraints on some types of measurement (for example using cookies online[3]), concerns include fairness, the privacy and confidentiality of personal data, and ethical treatment of both searchers and researchers.

In many industrial settings, as well as some research settings, privacy guarantees drastically limit what metrics can be used. Private data, such as email or corporate files, is an obvious instance: most providers will strictly limit access even if doing so denies themselves access to queries, documents, or metadata. Most offline metrics are impossible in this situation, as are most self-reports (without careful vetting to avoid exposing private data), and any process involving a third party. Evaluation has generally been limited to implicit feedback [51, 86], which requires a production-ready system; or public or synthetic corpora [6, 38], which may or may not resemble the real queries and documents. "Eyes-off" labelling may be possible with large language models [16, 85, 101], but this is not yet commonplace.

Ethical concerns extend not just to searchers, but also to researchers and contracted workers. Crowdsourcing, or even much in-house labelling, involves third parties so adds extra considerations. Payment rates must be appropriate, as must schedules. Especially in web search, there is also a real risk of exposing workers to undesirable queries or documents—at Bing we have established metrics for dealing with adult content and possible hate speech, but take care to recruit and train carefully and separately from other tasks.

4 VALIDITY

For a measurement to be useful, it needs to tell us something about the world; it must correlate with (or predict) some phenomenon of interest and generalise to new situations or as-yet unseen examples. This is *validity*. We consider three aspects of validity in particular: construct validity (are we measuring a real thing?), external validity (can we learn about new things?), and concurrent validity (does this measure look like others?).

4.1 Construct validity

Construct validity asks: are we measuring a real phenomenon? Does our metric faithfully represent the real construct? Here, our constructs are (or should be) the things we really care about: search quality, or task support, or some other phenomenon. This is often a hard problem; ensuring validity takes work [66]. For many important goals the objects we care about are difficult to measure, and proxies are poor. Metrics which are appealing are not always useful.

Harman [36] discusses choosing a metric, and rules for "relevance", with regard to a user model or a model of an information-seeking scenario. She suggests, for example, that for high-recall tasks MAP is appropriate (or MRR where a single document is sought), or that success at 1 or nDCG are valid choices for web search.

These recommendations are backed by a sense of which models are useful in which circumstance, but some search literature supports correlations between our common metrics and (slightly) higher-order phenomena. McDuff et al. [62] worked with recordings of search activity and facial expression recognition, and saw a correlation between simple online metrics and facial expressions—for example, an increase in long-dwell clicks was correlated with expressions of happiness, and queries that went on to be reformulated were associated with expressions of anger. Other groups have used questionnaires to show a correlation between offline metrics and (e.g.) searcher satisfaction. As one example of this approach, Chen et al. [13] show moderate correlation between metrics in the C/W/L/A family and final satisfaction on the THUIR1 data set, and use this correlation as a measure of their new metrics. This was popularised by Chen et al. [15] but has seen much use, with broadly good correlations [e.g. 14, 43, 92, 100]. Along similar lines, Azzopardi et al. [5] and Thomas et al. [84] saw correlations between offline metrics and online behaviours such as query reformulation and time on the SERP, and again used these correlations to argue for different forms of metrics.

These results should give us some faith in the validity of offline metrics, but represent only one small step towards predicting properties such as long-term satisfaction, utility, or revenue. It remains extremely hard to draw a line from an offline metric, over a collection of results, to the quality of a whole SERP let alone to anything "higher".

We also require that a metric not just represent a thing in the world, but that the thing be something we care about. Experimentation and machine learning are powerful optimisation tools, as seen in search systems from Facebook, Google, Netflix, Spotify and others, but it is easy to optimise for metrics which do not align with business goals. In general, where construct validity is less than perfect there are two risks: that a gain on our proxy metric is actually neutral (or negative) on the real goal, or that real gains are missed since the proxy metric is flat or shows a loss.

For example, Netflix tried deep learning in their recommender systems, optimising towards clicks. This predicted clicks well, but did not translate to better subscriber retention [33]: the problem of course is that the connection between the proxy metric and the business interest is "quite complicated and may even break in certain ranges" [82]. Spotify have noted a similar phenomena, that supporting immediate goals (by emphasising relevance) means that less-diverse music gets suggested. On the other hand, users who see more diverse music—and learn the catalogue—are more likely to stay subscribed, and become less passive. Optimising the usual short-term metrics would eventually hurt the business [3].

Metrics may also have a more complex relationship with the phenomenon of interest than we first anticipate. As one example, we have observed that as a search system improves, people will try harder queries and offline metrics will slip—the metric will fall even

[3] https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en

Dimension	Description	Examples
Social and legal—Should we measure this at all? In this way?		
	The extent to which a metric fits social, ethical, and legal constraints.	Legal compliance; ethical standards; privacy and confidentiality.
Validity—Are we measuring what we think we are measuring?		
	Whether the designed metric is faithful to our measurement goals: whether it measures what we expect it to measure.	Construct validity; external validity or generalisability; concurrent validity or agreement with existing metrics; inter-rater reliability.
Efficiency—Can we afford it?		
	The cost of collecting data.	Judging cost, and time; end-to-end time; unusable labels; overlap; scalability across languages and geography.
	Contribution to other running costs.	System development and maintenance cost, and time; technical expertise to maintain system; reliability; judge complaints; downstream impacts; difficulty of incremental improvements.
	Impact on constrained resources.	Need for expert judges.
Reliability, sensitivity—Can we get a reliable signal?		
	Ability to reliably detect change and to separate signal from noise.	Label distribution; statistical power; movements given controlled changes; meaningful gaps between systems.
Interpretability—Do we understand what the metric says?		
	Aspects such as simplicity/complexity, intuitiveness, and metric debuggability.	Debuggability of the systems under test; debuggability of the metric; complexity of the metric and underlying model(s).
Organisational—What happens when we report this metric?		
	Whether the metric leads to undesirable effects in the organisational setting.	Interactions with other components or measures. Consequences of treating a metric as a target.

Table 2: Summary of metric dimensions surveyed in this paper, and examples of properties a metric might or should have.

as people have more confidence in the system overall. As another, a change in usage may not signify a change in search quality but instead a change in some other part of the system (for example, a UI redesign). In this case, an increase in activity may signify that something has gone wrong somewhere else.

4.2 External validity

External validity is the extent to which conclusions from our metric generalise to other, unmeasured, cases. What does one experience tell us about anyone else's?

A standard approach is first to ask whether our samples are representative of some wider population. A representative sample is already often difficult, and in direct opposition to other desirable properties. For example, queries often follow a power-law distribution [69], so if we sample uniformly from a log then we will tend to get many examples of a few distinct queries. These popular queries tend to be for relatively easy navigational tasks, and therefore we will have little statistical power and little insight into other types of query. On the other hand, if we stratify or re-weight our sample then we are no longer measuring a representative workload. Similar arguments can be made for sampling searchers, languages, task types, or most other properties.

Similar sampling problems come from the labelling process itself. Background, cognitive style, geography, and personality all make a difference to relevance judgements [46, 47, 78], so a judge pool without realistic variation on any of these dimensions will not represent all our searchers. At the same time, if we use crowd judges then it may not be possible to collect reliable demographic data, and with any source it may not be possible to recruit the right mix of people.

Interactions between search components can also make it hard to generalise from one scenario to another. If we have two apparent improvements a and b, which lead to increases in a metric Δ_a and Δ_b, in general we cannot assume that the combination of improvements will increase the metric by $\Delta_a+\Delta_b$: interactions between the improvements could lead to something less than the sum of the parts, or (rarely) something better or even an overall degradation. As a trivial example, increasing the font size on a SERP and increasing the whitespace on a SERP might each make it easier to use but the combination—large type *and* lots of whitespace—might be unusable. Armstrong et al. [4] demonstrated these interactions with 2^6 combinations of "improvements" in the Indri search engine. Most were not simply additive, and some degraded overall effectiveness.

With any live system, changes can happen outside of controlled experiments. There might be new searchers, different query distributions, or new documents; or perhaps searcher behaviours and

preferences change with the system. This dynamism threatens generalisation, since conclusions drawn from our measurements at one time might not hold at any later time. One option is a held-out sample, isolated as far as possible from day-to-day changes, to act as a reference; however, by isolating a sample from changes, we lose validity over time. Harman and Buckley [37] demonstrate one way to manage this, by running current systems on earlier (frozen) data, but this is not plausible in cases such as web or enterprise search where we cannot freeze the entire input corpus.

Another approach, particularly with online metrics, is a long-term held-out sample: a group of searchers who are not exposed to any new features or experiments. Facebook report six-month holdouts [19], for example, and Pinterest report one year [28]. These holdouts let them measure the impact of a change over time, as searchers move from first experimenting with a feature to expecting or ignoring it; this also helps confirm that leading metrics (such as offline metrics) do correlate with long-term value.

4.3 Concurrent validity

Finally, we might consider concurrent validity: the extent to which two metrics, which purport to measure the same thing, do in fact agree. This criteria is often used in information retrieval: it is common to argue for a metric by establishing that it correlates well with another, more established, alternative. For example, bpref and the induced AP family were explicitly designed to cleave closely to AP [12, 94]. Other work has used correlation between metrics to group them, and to draw conclusions about which to use [e.g. 8, 30, 44, 87].

In general, good correlation with some earlier metric is desirable if that metric shows construct and external validity. For metrics like bpref and induced AP, which are motivated by cost and data availability, higher correlations are better, but for most work we argue instead that only moderate correlation is desirable. A metric which correlates poorly with some earlier, trusted, metric might simply be measuring the wrong thing; but a new metric which correlates very highly with an established one isn't measuring anything new.

5 EFFICIENCY

Producing any metric comes at some cost, so the efficiency of the process itself is a consideration. Salient aspects include up-front cost, time, and running costs. Retrieval itself, computing metrics, and statistical tests are normally cheap; the cost of labelling dominates.

5.1 Cost

The most obvious efficiency parameter is just: what is the dollar cost to get a measurement, a label, or a comparison?

Labelling has historically been expensive, with expert annotators involved as early as the Cranfield experiments [18]. The cost of annotators motivates the ongoing popularity of crowd workers, providing labels much more cheaply, although often at the expense of quality [1, 2, 7, 22]. The recent rise of labelling with large language models [16, 85, 101] offers further significant savings. Cheaper labelling, of course, gives better statistical properties for the same budget.

Even with a fixed cost per label, metric choices lead to different labelling requirements and therefore different costs. There is often a tradeoff between cost and sensitivity—more labels increase sensitivity, but also cost—but the design of a metric also makes a difference. An example comes from Sakai and Kando [77], who reran TREC evaluations with progressively fewer labels. Most metrics gave inconsistent conclusions as labels reduced. Those that were consistent—bpref and modified versions of AP, Q, and nDCG—could be run for a lower cost while drawing similar conclusions.

5.2 Time

Of course time also matters: both the time to get each label, and the time to run an experiment end-to-end and to draw a conclusion. This is important since our usual metrics, on- or off-line, are generally intended as "leading indicators" for slower-moving phenomena: they are valuable precisely because we can reach conclusions sooner [e.g. 26, 27]. End-to-end time also matters, naturally, if experimenters or others need data to commit to some decision.

Time per label and end-to-end time often correlate, but need not. Some measurement processes are quick to start and are fast per label, but have limited capacity: for example, a search expert doing their own labelling. Other processes are slow to start but can eventually produce labels rapidly: for example, a crowd may need time for initial training. End-to-end time can also be a property of the objects being measured, not the measurement process: for example, measuring rare events will naturally take longer than measuring common ones.

Industrial measurement also demands *scalability*; that the end-to-end time be consistent across larger or different workloads. A process which is fast with thousands of objects may be slow with hundreds of thousands (for example due to capacity constraints), or may slow down over time (for example if staff are depleted). Breakdowns in a measurement system can also cause backlogs. This can also be apparent if we want to measure a new aspect, or change direction: at Bing, latency has been a difficult problem when we add languages or regions, and hence need to recruit and train workers. In our experience, the more a metric is tied to the particular environment of its development, the more likely it is to be expensive in the long run.

5.3 Contribution to running costs

In a running system, unlike most research systems, measurement is ongoing and fits into a larger system of day-to-day work. Running cost is therefore a further consideration. Costs in this case are not just financial or time—although these are important—but include the effort of system development and maintenance; the need for internal or external expertise, generally in short supply; demand for computing, storage, or similar resources; reliability and the cost of downtime; reputational cost if crowd workers cannot complete the task, or are not compensated well; and the difficulty of changing and improving the metric itself. Any of these factors can contribute materially to long-term metrics efforts.

6 RELIABILITY AND SENSITIVITY

For a metric to be useful, it has to show us differences whenever—and only when—they truly exist. The metric, and associated processes, needs to reliably detect change and to separate signal from noise [25, 76].

One reason to measure is to demonstrate value (§1.1). Very often that reduces to a claim that system A is better than system B, and that claim is on the back of a null hypothesis test. We should therefore care about statistical properties, in particular the power of our test (linked

to the variability of our metric) and whether we can see meaningful gaps between systems (or discriminate between systems).

In offline metrics this depends in part on labelling, particularly the distribution of labels: for example, if we label mainly easy queries then many will score 100% and cannot any more distinguish between systems. For this reason we have found it useful, in scenarios where typical searches are simple searches, to stratify our sampling and focus more on hard cases. In all metrics this also depends in part on how we make aggregate scores: for example, if we use deeper metrics then we should be more sensitive to differences. Finally, it depends in part on how we combine scores to make decisions, or the kinds of statistical processes we use: some tests have more power than others [74, 81].

We would also like to know that the metric we produce reflects as far as possible the quality of the system, not just the system under a particular circumstance: that is, we want a metric to be stable if we change the coincidence of queries, documents, or timing. Distinguishing interesting from exogenous change is especially difficult given a corpus, a pool of searchers or judges, and a workload all shifting at the same time, but some experimental work has demonstrated the scale of the problem. Craswell et al. [20] bootstrapped data from MSMARCO, ranking systems by their effectiveness each time, and saw in some cases large changes as the query sample changed; a different sample can lead to different conclusions. Culpepper et al. [21] and Bailey et al. [8] report a similar effect due to the happenstance of query phrasing given a fixed information need.

A common pattern is that the distribution of a metric changes when we measure under different circumstances, but system orderings are much more stable. In this case we still believe we know which system is best, but can't conclude anything from the magnitude of the metric. Even system ordering can change, however, as we do things like change judges or change judging guidelines [7, 45], so caution is advised and reliability cannot be assumed.[4]

An unstable metric makes decisions hard. Kohavi et al. [53] discuss a metric where changes week over week were an order of magnitude larger than changes due to the experimental treatment; it would be very easy to be misled in such a case. We have seen similar effects at Bing: for a time we measured performance using different queries each day, so we could catch trends like sports events or happenings in the news. Unfortunately the day-by-day variance was such that (while useful for diagnostics) the metric became largely useless for reporting, and in later work we reduced the turnover.

At a minimum, we want *repeatability* if we measure the same configuration twice—i.e., if the objects being measured do not change, neither should the measurement. Even this is not always given: Thomas et al. [83] reported a case where a metric shifted unexpectedly from one day to the next. This was attributed to changes not in systems, queries, documents, or crowd workers, but to holidays influencing workers' schedules and mood. Capturing as much metadata as possible can help debug these problems as they occur.

[4] A standard countermeasure is to use at least two sets of data, e.g. a "test"/"train" split. If the whole data set is infrequently updated, however, there remains a risk of overfitting: overstating our performance on this particular set, possibly even to the point of making decisions which would be reversed on another set.

7 INTERPRETABILITY

A metric is a way to *measure* the quality of a system but also a way to *communicate* that quality, potentially to varied audiences including engineers, researchers, management, and searchers themselves. This communication is of course much easier if the metric is easy to interpret. With a metric which is interpretable—not just valid—we can identify losses in fine detail; categorise classes of problems; get to the root cause of problems; understand algorithms, their strengths, and their weaknesses; and find opportunities to improve both the search system and the metric itself.

In particular, an interpretable metric allows two kinds of debugging (and improvement). First, we can *debug the search system*. Given a poor result, we can use a mental model of that metric to understand where the system has failed, and start analysis or debugging from that point. Harman [36] similarly highlights a need for "better metrics and methodology for diagnostics".

We can also *debug the metric*, if we're sure the system is right (or wrong) and the metric says otherwise. Are the labels wrong?, the way they're combined?, something else? Understanding what the metric *should* do is obviously useful in this case.

All else equal, this suggests simpler metrics even at the loss of some validity. For example, adaptive metrics—those where the contribution of one item depends on the contributions of others—are less debuggable because of this interaction. As an extreme case, a perfectly relevant document contributes nothing to reciprocal rank if another relevant document appears earlier: no amount of examining this one document will explain the score. More complex metrics such as IFT [5], BPM [99], or session-level metrics [93] similarly make it harder to debug systems.

Wanting an interpretable metric points to simple maths and simple models, but also to separation of concerns. At Bing, our internal debugging tools show labels over each of several aspects, for each document, rather than a single score. This makes it simpler to (dis)agree with either the system or the metric, since it is easier to evaluate a single aspect than to evaluate several at once and do some form of weighting or combination. This suggests, for example, breaking "relevant" into things that are easier to reason about—perhaps topicality, recency, authority, or language—and labelling each.[5]

Any running search system is complex, likely with many teams working more or less independently on each of many parts. A good metric, or set of metrics, should therefore help with attribution—signalling which part of the whole system is performing well (or badly), or measuring each part for its contribution to the whole. An appropriate suite of metrics can be useful here, although it is possible to use a single end-to-end metric alongside techniques such as ANOVA or grid searches to attribute credit [21]. "Attribution mechanisms" have been studied in economics, for example to value the contribution of multiple ad exposures [9], but to our knowledge no similar techniques have been tested in IR.

Finally, a good metric will also help generate hypotheses. To make a causal claim ("X because Y") needs an experiment, designed or natural. That is, if all we have is a measurement (on any metric) then

[5] As a counter-point, we note that judging guidelines at Google describe various aspects but ask for a single combined score at the end.

we're able to say *how well* a system performs, but not *why*. As designers and decision-makers, we want to tell causal stories, so a good metric will let us suggest hypotheses for experimental validation.

An example comes from a web-based communication tool, where a key metric was daily feedback on a Likert-like scale, collected from users in situ. This metric was consistent for 200 days from its inception, then showed a sudden and consistent decline. This alone was not enough to say what caused the drop, but the daily measurements plus substantial metadata allowed one of us (Craswell) to quickly generate and test several hypotheses: perhaps a new feature was misbehaving, for example (testable by looking at the software versions reporting), or perhaps the workload had changed (testable by looking at anonymised interactions), or perhaps day 200 was itself special due to a bug in the logging code or some other accident (testable by looking at the volume and characteristics of the logs). That these hypotheses spring to mind so easily is a useful feature of the original metric and tools.

8 ORGANISATIONAL EFFECTS

Metrics have social lives, and any sufficiently influential metric will lead to changes in an organisation. In particular, a valid metric is an abstraction of something that matters, but once the abstraction is a target there is an inevitable effect: we will work to maximise the metric, with little or no attention to the slippery and complex phenomenon it proxies for. Islam and Greenwood [41] call this "performative capacity": the ability of metrics "to shape the world in their image and to bring new social realities into being".

8.1 Goodhart's law

A famous cartoon appearing in Krokodil magazine has the foreman of a Soviet nail factory assigned to produce a certain weight of nails. The factory produces one enormous nail; when told it's useless, he replies "it's nothing. It's important that we fulfilled the plan immediately"[6]. The joke of course is that if we are rewarded for producing a certain mass of nails, we will, even if it doesn't help any real carpenters. (Another variant has management switch to rewarding the number of nails—and getting lots of small ones.)

Real-world examples of this phenomena include a bounty on cobras in Delhi, which (rather than reduce snakes) saw cobras bred for income. When the bounty was dropped, breeding snakes were released, making matters worse [59]. Similar missteps attended programmes for rats in Hanoi in 1903 [88] and, 100 years later, for opium poppies in Afghanistan [91]; we might also consider payments for health interventions (not preventative health), or many other so-called "perverse" incentives.

These can be seen as instances of Goodhart's law [17, 34], paraphrased by Hoskin: "when a measure becomes a target, it becomes a bad measure" [40]. The danger is that our true goal T is hard to measure; so we have a surrogate metric S; and now if our decision-making targets S, then we will see some bad outcomes on T.

Examples include T= research excellence, S = number of papers published, with the outcome of many minimum-quality papers; or T= good education in schools, S = standardised test results, and the outcome of teaching to the test. Many further examples suggest themselves in search:

- To maximise *revenue*, we might insert more advertisements.
- To maximise *clicks*, we might promote clickbait.
- To maximise *time on page*, we might promote unreadably bad page design.
- Maximising *usage* might mean some other part of the product is broken (leading to search as a last resort), or that search is so poor that searchers need several attempts.
- To maximise *mean utility* across results, we might never take risks—only showing the most obviously relevant documents.
- To maximise *total utility*, however, we make take silly risks and include anything with any chance of being even slightly relevant.
- By maximising *checkout rate* or *purchases per session* in an online shop, we could miss chances to grow the number of shoppers and total revenue.

This is not just a hypothetical concern. By about 2016, YouTube's and Facebook's experimentation and machine learning was so powerful that, aiming at "engagement", their systems tended to radicalise and infuriate users: maximising engagement but at considerable cost [71]. Facebook, and YouTube's parent company Google, admitted to this misalignment in mid 2017 but only after considerable public and regulatory pressure [58, 73].

8.2 Lock-in

Around the year 2010, the part of Microsoft which included Bing was losing about US$2B per year [80]. With very little market share and no real revenue, the organisation was measuring search quality via nDCG [42] using crowd-sourced labels. Teams were committed to this metric, as far up as board level.[7]

At times, that would lead to clearly bad decisions. Figure 1 illustrates, with data from the time, the top results for the query [mahjong]. Human ratings ("HRS") were preferring the Wikipedia entry to the online game—this accorded with Bing's guidelines, as Wikipedia is generally broad and reliable as well as being a well-known brand.[8] Searchers, however, weren't looking for a trusted source on the game's rules or history: click data ("CTR", clickthrough rate) hugely favoured the online game.

Examples such as this, where quality labels were misleading even though they were correct, were relatively easy to come by. With everyone up to the board invested in nDCG, however, the hard part was in agreeing to take a loss on the metric that mattered most, and look at other metrics as well.

This is an example of what Lanier [56] calls *lock-in*: a successful design constraining later possibilities, entirely because of its initial success. Other examples in IR evaluation might include using MAP, despite other metrics being more appropriate in many scenarios; or indeed the entire TREC model of third-party judges, pooling, and small- to medium-scale query sets. It is now hard to imagine a system being accepted as "better" without a t test over TREC offline metrics, even in cases where better options exist.

[6] "Это пустяки . . . Мы сразу выполнили план по гвоздям". We have not been able to find the issue where this cartoon first appeared.

[7] We are indebted to Bill Ramsey for this anecdote.

[8] Kazai et al. [45] reported a similar bias towards Wikipedia and other popular sites, when labelling with crowd judges.

Figure 1: Bing results for the query [mahjong], mid July 2011. A commitment to "HRS" ratings meant a more-popular site (mahjonged.com) was ranked lower than a less-popular (Wikipedia).

8.3 Tensions between metrics

One obvious mitigation is to report on, and base decisions on, more than one metric: for example, by measuring both clicks and abandonment we may be able to see when we are maximising clicks at the expense of relevance. Having multiple targets is harder both for machine learning and for human decisions, but has the advantage of forcing us to explicitly consider the tensions and tradeoffs.

Often these are tradeoffs between different teams: for example, a ranking team may be maximising clicks (as an indication of relevant results) while a "rich results" team may be maximising abandonment (as an indication that good information is on the SERP and no clicks are needed). Similar tensions have surfaced at Bing, where a desire to maximise revenue has, from time to time, led to more ads and lower overall quality; at the same time, maximising engagement with "organic" results at the expense of ad clicks would send the company broke. This needs careful management, but in our experience this is preferable to any monolithic—and uninterpretable—single metric.

9 METRICS PRACTICE AND RESEARCH

Search metrics are tools for communication and decision-making by humans (and in some contexts by machines). This makes it important to carefully consider the properties of metrics, and to be deliberate about metric design and adoption.

9.1 Metrics in practice

The list above is daunting, and it is hard to know how well a metric satisfies these desiderata. It is useful to have a process for new metrics, to check against this list and to understand the trade-offs. It is also useful to call on "metametrics" to measure the metrics themselves [25, 35, 57].

There are few examples of metrics development processes in the literature. Mustafa and Khan [65] have a process ("qMDF") for developing measures of software quality; however, we are not aware of many similar attempts in IR or related fields.

At Bing, we have a process and set of checks for any new offline metric. It is based on two data sets: a corpus of experiments, and a set of high-quality "gold" labels. First, over time it is possible to collect a *corpus of treatments*—new rankers, UX elements, or other changes—where we are reasonably sure of the outcome: for example, we may know that one treatment led to higher revenue, or lower usage. We expect our metrics to show a corresponding change. Second, we also use a *set of labels from real searches*. They are gathered from several sources: from employees in the context of their usual searching, from contractors who are paid to look through their search history, and from a "feedback" button on the Bing search results page. This data is provided at, or close to, the time of the seacher's actual need; by the searcher themselves; in the context of a full search session; and are reviewed by Bing employees. The labels are very reliable, and can serve as "gold" references [7] to test metrics.

We can now test any new metric against the desiderata discussed above. Our checklist includes whether we should measure at all; how well the metric correlates with other phenomena of interest or with existing metrics; whether the metric correctly identifies cases in our corpus and gold labels; how closely the metric is tied to a particular sort of sample; how stable the metric is day to day; cost, throughput, and latency; and how the metric interacts with other metrics, and with the larger team's incentives.

We have a choice what is in our test sets, so can use them to express Bing policy. For example, if we want to be sure our metrics distinguish spam, we can include spam and non-spam examples. This helps us ensure validity where it is most important. Similarly, we can emphasise metric sensitivity by including examples of borderline results, and as metrics improve we add more "hard" cases.

9.2 Recommendations for research

The research literature has long considered problems of *reliability and sensitivity*, some forms of *validity*, and *efficiency*. In research settings metrics are used to communicate results, but less commonly to debug or develop systems, and the *interpretability* of a metric has not been a prominent concern (we will note, however, a metric that helps to create hypotheses is likely to be more productive). Other aspects have received less attention.

Standard collections for offline work, and coverage by review boards for lab studies or online work, mean questions of *social and legal constraints* are not often considered in IR literature. The choices we make here have social implications nevertheless. For example, we tune many systems on the same data, meaning we privilege the small number of people providing topics and assessments; we work mostly in English; we do not normally report metrics for diversity; and we choose not to work on certain types of data. There is a large body of work on metrics for fairness, but little on the fairness of metrics, and this deserves more consideration.

Questions of metric *validity*, especially construct validity, are evergreen in the literature. Increasingly sophisticated simulations, including the ability of large language models to synthesise queries or conversations, need increasingly sophisticated tests if we are to be sure our simulated "searchers" are useful. This will continue to be key for any metric.

Finally, *organisational effects* are real even in a small group. It is worth considering how can we run a workshop, or a shared task, without locking teams in to one way of measuring progress.

ACKNOWLEDGMENTS

We are grateful to the many colleagues who have helped our understanding of metrics and measurement.

REFERENCES

[1] Omar Alonso. 2022. *The practice of crowdsourcing*. Springer Nature.

[2] Omar Alonso and Gary Marchionini. 2019. *The practice of crowdsourcing*. Morgan & Claypool Publishers.

[3] Ashton Anderson, Lucas Maystre, Rishabh Mehrotra, Ian Anderson, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on Spotify. In *Proceedings of the International Conference on World Wide Web*.

[4] Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 601–610.

[5] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: An information foraging based measure. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 605–614.

[6] Peter Bailey, Nick Craswell, Ian Soboroff, and Arjen P de Vries. 2007. The CSIRO enterprise search test collection. *SIGIR Forum* 41, 2 (2007), 42–45.

[7] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 667–674.

[8] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 625–634.

[9] Ron Berman. 2018. Beyond the last touch: Attribution in online advertising. *Marketing Science* 37, 5 (2018), 771–792.

[10] Nolwenn Bernard and Krisztian Balog. 2023. A systematic review of fairness, accountability, transparency and ethics in information retrieval. *Comput. Surveys* (Dec. 2023). To appear.

[11] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam D I Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489 (2012), 295–298.

[12] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25–32.

[13] Nuo Chen, Jiquin Liu, and Tetsuya Sakai. 2023. A reference-dependent model for web search evaluation: Understanding and measuring the experience of boundedly rational users. In *Proceedings of the International Conference on World Wide Web*.

[14] Nuo Chen, Fan Zhang, and Tetsuya Sakai. 2022. Constructing better evaluation metrics by incorporating the anchoring effect into the user model. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2709–2714.

[15] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[16] Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluation? arXiv:2305.01937v1 [cs.CL]

[17] K. Alec Chrystal and Paul D. Mizen. 2001. Goodhart's law: Its origins, meaning and implications for monetary policy. Prepared for the Festschrift in honour of Charles Goodhart.

[18] Cyril W Cleverdon, Jack Mills, and E Michael Keen. 1966. Factors determining the performance of indexing systems, volume 1: Design. Aslib Cranfield Research Project.

[19] Alex Coleman. 2021. Introducing holdouts. https://blog.statsig.com/introducing-holdouts-4bcfc1821d1c. Accessed January 2024.

[20] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. MSMARCO: Benchmarking ranking models in the large-data regime. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1566–1576.

[21] J Shane Culpepper, Guglielmo FAggioli, Nicola Ferro, and Oren Kurland. 2021. Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems* 40, 1, Article 19 (2021).

[22] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *Comput. Surveys* 51, 1, Article 7 (Jan. 2018).

[23] Alex Deng and Xiaolin Shi. 2016. Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 77–86.

[24] Djellel Difallah and Alessandro Checco. 2021. Aggregation techniques in crowdsourcing: Multiple choice questions and beyond. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 4842—-844.

[25] Pavel Dmitriev and Xian Wu. 2016. Measuring metrics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 429–437.

[26] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the International Conference on World Wide Web*. 256–266.

[27] Georges Dupret and Mounia Lalmas. 2013. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 173–182.

[28] John Egan. 2015. Long-term impact of badging. https://jwegan.com/growth-hacking/long-term-impact-badging/. Accessed January 2024.

[29] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS* 112, 33 (2015).

[30] Nicola Ferro. 2017. What does affect the correlation among evaluation measures? *ACM Transactions on Information Systems*, Article 19 (Aug. 2017).

[31] Nicola Ferro and Mark Sanderson. 2022. How do you test a test? A multifaceted examination of significance tests. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 280–288.

[32] Peter B. Golbus, Imed Zitouni, Jin Young Kim, Ahmed Hassan, and Fernando Diaz. 2014. Contextual and dimensional relevance judgments for reusable SERP-level evaluation. In *Proceedings of the International Conference on World Wide Web*. 131–142.

[33] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems* 6, 4 (Dec. 2015).

[34] Charles A E Goodhart. 1975. Problems of monetary management: The UK experience. In *Papers in Monetary Economics*. Vol. 1. Reserve Bank of Australia.

[35] Somit Gupta and Widad Machmouchi. 2022. STEDII properties of a good metric. https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/stedii-properties-of-a-good-metric/. Accessed January 2024.

[36] Donna Harman. 2011. *Information Retrieval Evaluation*. Number 19 in Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan and Claypool.

[37] Donna Harman and Chris Buckley. 2009. Overview of the reliable information access workshop. *Information Retrieval Journal* 12 (2009), 615–641.

[38] David Hawking, Bodo Billerbeck, Paul Thomas, and Nick Craswell. 2020. *Simulating information retrieval test collections*. Synthesis lectures on information concepts, retrieval, and services, Vol. 71. Morgan and Claypool.

[39] Alex Hern. 2014. OKCupid: we experiment on users. Everyone does. *The Guardian* (29 July 2014). https://www.theguardian.com/technology/2014/jul/29/okcupid-experiment-human-beings-dating

[40] Keith Hoskin. 1996. The 'awful' idea of accountability: Inscribing people into the measurement of objects. In *Accountability: Power, ethos and technologies of managing*, R Munro and J Mouritsen (Eds.). International Thompson Business Press, London.

[41] Gazi Islam and Michelle Greenwood. 2022. The metrics of ethics and the ethics of metrics. *J. Business Ethics* 175 (2022).

[42] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[43] Jiepu Jiang and James Allan. 2016. Correlation between system and user metrics in a session. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 285–288.

[44] Timothy Jones, Paul Thomas, Falk Scholer, and Mark Sanderson. 2015. Features of disagreement between retrieval effectiveness measures. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 847–850.

[45] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and S M M Tahaghoghi. 2012. An analysis of systematic judging errors in information retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 105–114.

[46] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 2583–2586.

[47] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval Journal* 16 (2013), 138–178.

[48] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1–2 (2009), 1–224.

[49] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum* 37, 2 (2003), 18–28.

[50] Jinyoung Kim, Gabriella Kazai, and Imed Zitouni. 2013. Relevance dimensions in preference-based IR evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 913–916.

[51] Jin Young Kim, Nick Craswell, Susan Dumais, Filip Radlinski, and Fang Liu. 2017. Understanding and modeling success in email search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 265–274.

[52] Jin Young Kim, Jaime Teevan, and Nick Craswell. 2018. Explicit in situ user feedback for web search results. In *Proceedings of the International ACM SIGIR*

Conference on Research and Development in Information Retrieval. 829–832.
[53] Ronny Kohavi, Thomas Crook, Roger Longbotham, Brian Frasca, Randy Henne, Juan Lavista Ferres, Tamir Melamed, and Juan M. Lavista Ferres. 2009. Online experimentation at Microsoft. https://www.microsoft.com/en-us/research/publication/online-experimentation-at-microsoft/. Accessed January 2024.
[54] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18 (2009), 140–181.
[55] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790.
[56] Jaron Lanier. 2010. *You are not a gadget: A manifesto.* Alfred A Knopf, New York.
[57] Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-evaluation of conversational search evaluation metrics. *ACM Transactions on Information Systems* 39, 4, Article 52 (Sept. 2021), 42 pages.
[58] Natasha Lomas. 2017. Google to ramp up AI efforts to ID extremism on YouTube. https://techcrunch.com/2017/06/19/google-to-ramp-up-ai-efforts-to-id-extremism-on-youtube/.
[59] David S Lucas and Caleb S Fuller. 2018. Bounties, grants, and market-making entrepreneurship. *The Independent Review* 22, 4 (2018), 507–528.
[60] Widad Machmouchi, Ahmed Hassan Awadallah, Imed Zitouni, and Georg Buscher. 2017. Beyond success rate: Utility as a search quality metric for online experiments. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 757–765.
[61] Widad Machmouchi and Georg Buscher. 2016. Principles for the design of online A/B metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 589–590.
[62] Daniel McDuff, Paul Thomas, Nick Craswell, Kael Rowan, and Mary Czerwinski. 2021. Do affective cues validate behavioural metrics for search?. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[63] Alistair Moffat. 2013. Seven numeric properties of effectiveness metrics. In *Proceedings of the Asia Information Retrieval Societies*.
[64] Alistair Moffat. 2022. Batch evaluation metrics in information retrieval: Measures, scales, and meaning. arXiv:2207.03103 [cs.IR]
[65] K Mustafa and RA Khan. 2005. Quality metric development framework (qMDF). *Journal of Computer Science* 1, 3 (2005), 437–444.
[66] Heather L O'Brien and Elaine G Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the Association for Information Science and Technology* 61 (2010). Issue 1.
[67] Sanghee Oh and Barbara M Wildemuth. 2017. Think-aloud protocols. In *Applications of social research methods to questions in information and library science* (2 ed.), Barbara M Wildemuth (Ed.). Libraries Unlimited, Santa Barbara, California, Chapter 21, 198–208.
[68] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D. Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W. Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N. Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D. Ekstrand, Preethi Lahoti, and Toshihiro Kamishima. 2021. FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. *ACM SIGIR Forum* 53, 2 (March 2021), 20–43.
[69] Casper Petersen, Jakob Grue Simonsen, and Christina Lioma. 2016. Power law distributions in information retrieval. *ACM Transactions on Information Systems* 34, 2, Article 8 (Feb. 2016).
[70] Bahareh Rahmanian and Joseph G. Davis. 2014. User interface design for crowdsourcing systems. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 405–408.
[71] Manoel Horta Ribeiro, Robert West, Raphael Ottoni, Virgílio A. F. Almeida, and Wagner Meira Jr. 2021. Auditing radicalization pathways on YouTube. arXiv:1908.08313 [cs.CY]
[72] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. *Information Processing and Management* 58, 6 (Nov. 2021).
[73] Kevin Roose. 2020. Rabbit Hole. The New York Times: https://www.nytimes.com/column/rabbit-hole.
[74] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 71–78.
[75] Tetsuya Sakai. 2014. Statistical reform in information retrieval? *ACM SIGIR Forum* 48, 1 (2014), 3–12.
[76] Tetsuya Sakai. 2017. The probability that your hypothesis is correct, credible intervals, and effect sizes for IR evaluation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25–34.
[77] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval Journal* 11 (2008), 447–470.
[78] Parnia Samimi and Sri Devi Ravana. 2016. Effect of cognitive ability on reliability of crowdsourced relevance judgments. In *Proceedings of the International Conference on Information Retrieval and Knowledge Management*. 107–112.
[79] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375.
[80] Barry Schwarz. 2010. Microsoft records their Q1 2010 earnings, Bing takes loss again. Search Engine Land: https://searchengineland.com/microsoft-records-their-q1-2010-earnings-bing-takes-loss-again-54281.
[81] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 623–632.
[82] Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. 2021. Deep learning for recommender systems: A Netflix case study. *AI Magazine* 42, 3 (Nov. 2021), 7–18.
[83] Paul Thomas, Gabriella Kazai, Ryen White, and Nick Craswell. 2022. The crowd is made of people: Observations from large-scale crowd labelling. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 25–35.
[84] Paul Thomas, Alistair Moffat, Peter Bailey, Falk Scholer, and Nick Craswell. 2018. Better effectiveness metrics for SERPs, cards, and rankings. In *Proceedings of the Australasian Document Computing Symposium*.
[85] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.
[86] Pertti Vakkari, Michael Völske, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Modeling the usefulness of search results as measured by information use. *Information Processing and Management* 56, 3 (2019), 879–894.
[87] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Information Retrieval Journal* (2020), 411–448.
[88] Michael G Vann. 2003. Of rats, rice, and race: The great Hanoi rat massacre, an episode in French colonial history. *French Colonial History* 4 (2003), 191–204.
[89] Ellen M. Voorhees and Donna K. Harman (Eds.). 2005. *TREC: Experiment and evaluation in information retrieval.* The MIT Press, Cambridge, MA.
[90] Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. 2023. To aggregate or not? Learning with separate noisy labels. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2523–2535.
[91] Craig Whitlock, Leslie Shapiro, and Armand Emamdjomeh. 2019. Mohammed Ehsan Zia, lessons learned interview. The Washington Post: https://www.washingtonpost.com/graphics/2019/investigations/afghanistan-papers/documents-database/?tid=a_inl_manual&document=background_ll_04_xx4_04122016.
[92] Alfan Farizki Wicaksono and Alistair Moffat. 2020. Metrics, user models, and satisfaction. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 654–662.
[93] Alfan Farizki Wicaksono and Alistair Moffat. 2021. Modeling search and session effectiveness. *Information Processing and Management* 58 (2021).
[94] Emine Yilmaz and Javed A Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 102–111.
[95] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and effort: An analysis of document utility. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 91–100.
[96] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting civil discourse through search engine diversity. *Social Science Computer Review* 32, 2 (2014), 145–154.
[97] Eva Zangerle and Christine Bauer. 2022. Evaluating recommender systems: Survey and Framework. *Comput. Surveys* (2022).
[98] ChengXiang Zhai. 2022. Information retrieval evaluation as search simulation. Talk presented at the NTCIR-16 conference.
[99] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 425–434.
[100] Fan Zhang, Jiaxin Mao, Yiqun Liu, Xiaohui Xie, Weizhi Ma, Min Zhang, and Shaoping Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 379–388.
[101] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. arXiv:2310.14122 [cs.IR]