The Crowd is Made of People

Observations from Large-Scale Crowd Labelling

Paul Thomas
pathom@microsoft.com
Microsoft
Canberra, Australia

Gabriella Kazai
gkazai@microsoft.com
Microsoft
Cambridge, UK

Ryen W. White
ryenw@microsoft.com
Microsoft Research
Redmond, WA, USA

Nick Craswell
nickcr@microsoft.com
Microsoft
Bellevue, WA, USA

ABSTRACT

Like many other researchers, at Microsoft Bing we use external "crowd" judges to label results from a search engine—especially, although not exclusively, to obtain relevance labels for offline evaluation in the Cranfield tradition. Crowdsourced labels are relatively cheap, and hence very popular, but are prone to disagreements, spam, and various biases which appear to be unexplained "noise" or "error". In this paper, we provide examples of problems we have encountered running crowd labelling at large scale and around the globe, for search evaluation in particular. We demonstrate effects due to the time of day and day of week that a label is given; fatigue; anchoring; exposure; left-side bias; task switching; and simple disagreement between judges. Rather than simple "error", these effects are consistent with well-known physiological and cognitive factors. "The crowd" is not some abstract machinery, but is made of people. Human factors that affect people's judgement behaviour must be considered when designing research evaluations and in interpreting evaluation metrics.

CCS CONCEPTS

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

KEYWORDS

Crowdsourcing, Quality control, Cognitive biases

ACM Reference Format:
Paul Thomas, Gabriella Kazai, Ryen W. White, and Nick Craswell. 2022. The Crowd is Made of People: Observations from Large-Scale Crowd Labelling. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3498366.3505815

1 INTRODUCTION

When evaluating an information retrieval (IR) system, and particularly a ranking algorithm, it is almost universal to use relevance labels: judgements of how good a given document is for a given information need or query. These are aggregated with metrics such as the classic weighted-precision family which includes mean average precision, mean reciprocal rank, discounted cumulative gain (DCG), rank-biased precision (RBP), and others [50].

The labels in turn are commonly sourced not from searchers in situ but from third-party workers—so-called "judges"—who stand in for IR system users. These judges may be individual researchers or expert assessors [5], but increasingly are untrained strangers, operating at arms' length, via crowdsourcing platforms [15, 39, 43, 44, 49].

Crowdsourcing can drastically reduce the time and cost to label a large test collection [1, 18]. However, the reduced oversight means that the resulting labels may be lower quality (noisier) [2, 14, 35]. Some of this can be explained by buggy instruments or scales, difficult tasks, or bad actors amongst the judges; but, in our experience, a good proportion of data quality issues are due to human factors.

At Microsoft Bing, we often note these issues in crowd data. Past work has demonstrated relevance labels can be affected by cognitive biases or style [17, 30, 53, 55], work patterns [28], or existing beliefs [32], but this work has typically been at relatively small scale; or has induced the effect in question, not observed it "in the wild".

In this paper, we use hundreds of millions of labels, collected from hundreds of thousands of workers as a routine part of search engine development. With this we can demonstrate the impact of these human factors on the relevance judgement process, including some relatively large effects which, to our knowledge, have not been explored before in IR evaluation.

Before diving in to explore some specific factors and the impact on relevance judgement, let us begin with an example to help us frame and motivate this research.

Motivating example: A sudden metric change

In April 2021, we (authors and colleagues) at Bing noticed a sudden change in the metrics for a web search component. This component is evaluated in part by RBP-based scores [45], calculated daily over tens of thousands of judge-submitted labels. There is of course variation from day to day, as the ranking algorithm and the web evolve over time; but this variation is typically about 0.1 points on

CHIIR '22, March 14–18, 2022, Regensburg, Germany

ACM ISBN 978-1-4503-9186-3/22/03…$15.00
https://doi.org/10.1145/3498366.3505815

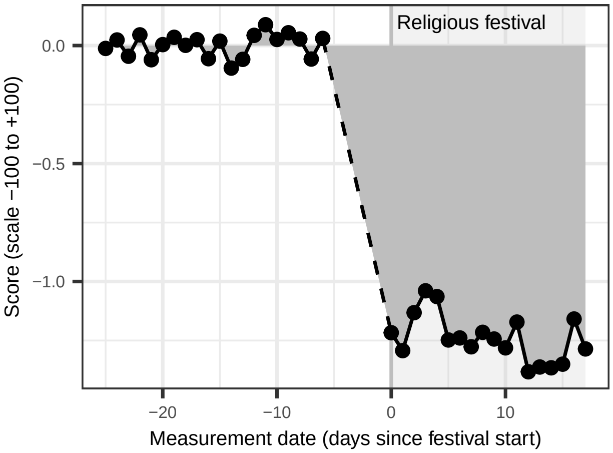


Figure 1: A sudden drop in an evaluation metric, possibly coincident with the start of a religious festival. The drop was an order of magnitude more than usually observed, but was not seen on later re-labelling. We also observed that the scores returned to normal after the festival concluded (not depicted in the figure).

our 100-point scale. However, in this case we observed close to a 1.5-point change almost overnight (Figure 1).

Close investigation showed that the change was due to crowd labels collected across a few days, and especially from judges based in a small number of countries. There had been no change to the judging system in this time, nor to the workload; and we were not aware of any changes to the ranking algorithm itself to warrant such a large change in score.

After collecting a further set of labels, the scores reverted to normal, so we are confident that the shift was "just noise"—albeit noise on a scale much larger than any true effect we would expect to see. Without recourse to interviewing the judges themselves, we will never know the true cause for the shift in labels. One theory is plausible, however: a major religious festival started at the time scores dropped, and the biggest changes were mostly in countries with significant populations observing the festival. Without specifying the religion we note that religious festivals can lead to changes in daily routines due to observances such as holidays, ceremonies and fasting, changes which are known to influence decision-making (although not necessarily at the cost of accuracy) [9, 24]. If we had accepted the metrics as they stood, or did not have enough experience to realise that they were unlikely, we could have been misled about a web-scale experiment because some of our judges were tired or hungry while labelling.

This anecdote serves to illustrate a more general point, which we rediscover regularly: when we see "noise" in crowd labels, it need not mean that the crowd is especially erratic or unpredictable. Rather, there might be some effect due to physiological factors such as tiredness or hunger, or due to well-understood cognitive biases and shortcuts. In this paper, we illustrate several examples we have seen in our extensive experiences with crowd labelling, discuss the effect on large-scale IR experimentation, and offer some advice. We hope this acts as a useful reminder that "the crowd" is not some imperfect peripheral, but is just a group of people we have not met.

In this paper, we present evidence of non-trivial effects on relevance estimation. These effects are due to biases, shortcuts, and time. They are certainly not the only effects at play, but were selected as they are well-attested in the literature, we believed they would be visible in our labels, and we believed the effects would be large enough to cause a difference in large-scale, production metrics. Our contributions are to (1) demonstrate these effects not in the laboratory but in large-scale existing data from ongoing evaluations of the Bing search engine, with trained and tested workers; (2) quantify the scale of these effects in a production setting; (3) illustrate some interactions amongst these effects; and (4) outline effects which have not to our knowledge been studied in an IR setting, including time-of-day, left-right, and mere-exposure biases. Note that given the differences in data and setup for each of the effects studied, we do not have an overarching methods section and instead describe the experimental methods for each effect separately.

2 JUDGES' LABELS VARY OVER TIME

As a first demonstration, we consider some temporal effects: that is, we look at whether and how the time that a document is labelled (time of day or day of week) influences the label that is given.

In this section, we base our analysis on labels of overall quality, each applied to a web document in the context of a given query. The data comes from a single "application" hosted in our internal crowdsourcing tool, the Universal Human Relevance System (UHRS), which operates similarly to Amazon Mechanical Turk.

In this application, judges scored documents on a five-point ordinal scale (bad, fair, good, excellent, perfect), based on attributes such as authority, recency, and topicality. We transform these linearly onto a numeric scale running 0 (bad) to 100 (perfect). The data comes from a long-running measurement system, with various uses, so we have taken only labels which we know came from production-quality ranking algorithms. We have several hundred thousand such labels, for regions and languages around the world.

If judges were acting as noisy conduits of relevance (or authority, or readability, or whatever other attribute), then as long as a document does not "go stale" it should not matter when labelling happens: a document which is (e.g.) "fair" will be "fair" whether it is viewed at 1am Tuesday or 2pm Wednesday. What we see, however, is substantial and non-random deviation in labels due to time.

2.1 Time of day

Figure 2 illustrates this with scores given to URLs in our crowd application. Scores here range from 0 to 100, and we aggregate these scores by query and by ranking algorithm to help determine which ranking algorithm is better, or whether there is a change in search effectiveness. Figure 2 shows, for each hour of the day, the mean score given by judges that hour (judge local time) compared to the mean score across the entire set. A score of zero denotes no difference from the overall mean. For example, at 2pm local time scores are almost 2 points higher than average. Each hour represents around 10,000 scores, for several hundred judges in almost 20 combinations of language and region, with more than 100 judges active each hour. We also observe these differences in

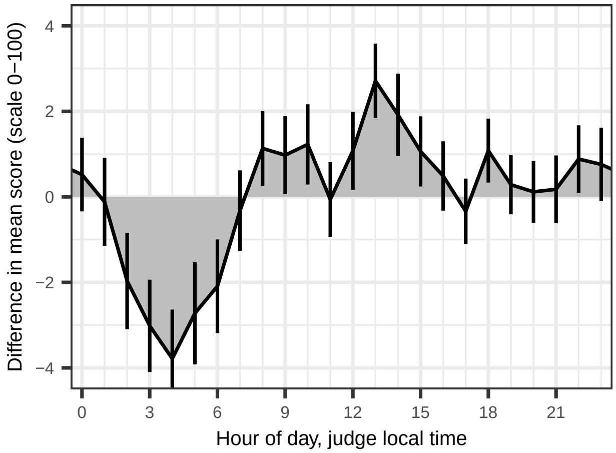


Figure 2: Differences in URL score (100-point scale) by judge local time. Judges working late at night (2–6am) give much lower scores; judges working in the morning and early afternoon give higher scores. Error bars denote ±1 SE.

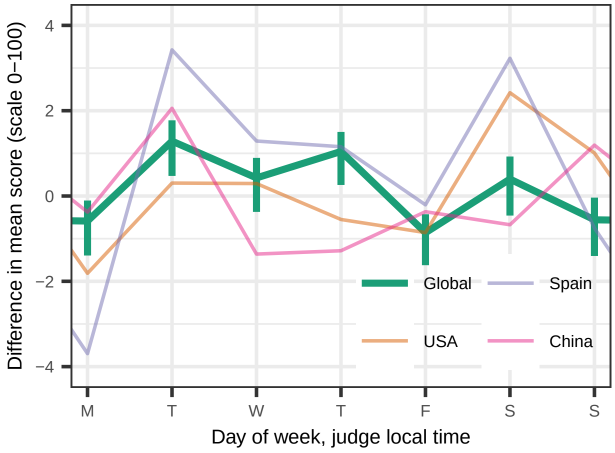


Figure 3: Differences in URL score (100-point scale) by judge local day of week. Judges working earlier in the week generally give higher scores than judges working later. We have around 20 regions and languages that comprise the "Global" line, so thin lines are examples only. Error bars denote ±1 SE.

a linear mixed-effects model, treating judge identity as a random term (t(295k) ranging to -5.5, $p \ll 0.001$)[1]: this indicates that the per-hour effect is not just a result of different judges being online at different times of the day.

When we compare different iterations of a search engine—for example, the production engine against an engine with a slightly improved ranking algorithm—we generally observe differences of at most a few points on this 100-point scale. The effect of judging URLs from one engine in the early morning, the other at lunchtime, may completely dominate the effect we want to see. Ideally, we could arrange for URLs from the "old" engine to be scored at 3am or 4am, and URLs from the "new" engine to be scored a few hours later at 12pm to 3pm, for a convenient head start. This is easier than it sounds: in fact it could happen by accident, as a baseline system may be scored some time before any experimental system(s). This may happen, for example, when experimental systems return previously-unlabelled documents, and we get new labels accordingly. Depending how often the baseline is updated, the difference in this case could be hours, days, or (in the case of long-running benchmarks such as TREC) even years.

One possible explanation for this swing in scores is a connection with patterns of mood across the course of a day. For example, Hernandez et al. [29] observed more smiling during the day but markedly less for those awake overnight; Golder and Macy [25] observed more positive affect in Twitter posts in a similar shape, although higher during the evening. There is in turn evidence that mood contributes to label quality and worker engagement in relevance labelling [23, 47], and that mood influences judgements and ratings in other fields, even with highly-trained workers [20].

Other factors may include fatigue, which we discuss below.

2.2 Day of the week

If we run evaluations with a daily cadence, and we know where our judges live, it should be more or less possible to release crowd work so that labels are being provided at the same judge-local time each day. However, we also see strong effects due to the day of the week (Figure 3). Again we plot the change in mean score given by judges on each day, in the judge's timezone. Here each day represents about 50,000 scores, with several hundred judges active each day and approximately the same volume of scores assigned each day across the week. In this case, we could get up to a two-point (dis)advantage just due to allocating different URLs on different days. Again, these differences are confirmed by a mixed-effects model (t(295k) ranging to 10.61, $p \ll 0.001$).

The thicker line in Figure 3 shows aggregate data, across the 20 global language-region combinations in which we had judgement data. The effect is, however, strongly mediated by the region a judge is from. Some examples are plotted as thinner lines. For example, judges in the USA give out sharply higher ratings over the weekend, while judges in China are less prone to this; judges in Spain give very low ratings on Mondays. Recent studies on mood have shown variations across different days and different countries [57].

2.3 Judge characteristics

We note that the judges whose data we use here were well-trained and were subject to ongoing audits. They were paid by the hour, not by the label, so there is little incentive to work fast at the expense of accuracy. Judges also had to read tens of pages of examples and guidelines, then pass an initial qualifying test, and were audited on an ongoing basis by comparing their scores to those of their peers and to those of specialist auditors. We expect that these checks will reduce exogenous influences such as time of day. This means that a less-trained worker, with less quality control and more incentive to

[1] Modelling used the lme4 and lmerTest packages in the R statistical software [7, 37, 48], versions 1.1-23, 3.1-2, and 4.0.2.

work fast than to work carefully, will in all likelihood experience *larger* effects than those discussed here.

3 JUDGES GET FATIGUED

We also observe an effect due to the time that judges have spent on their task. We define a "session" as a string of labels provided by a single judge, with no more than a one-hour break between items: "sessions" are delimited by a break of more than an hour[2]. Inside a session, a judge will see a variety of queries and web documents, and these will be randomised (at least amongst those available at the time the judgements are performed).

Using a mixed-effect model as above, we observe a small effect due to either the time in the session (1.09 points per hour, on our 100-point scale; $t(295\text{k}) = 12.43, p \ll 0.01$) or the number of scores already given in the session (0.013 points per judgement, $t(600\text{k}) = 32.90, p \ll 0.01$). We also observe a tendency to less variation in labels (decreasing difference between consecutive scores, 0.016 points per label, $t(588\text{k}) = -46.13, p \ll 0.01$).

These two effects mean that as time goes on, judges have a slight tendency to converge on good scores regardless of the query or document. This may be due to the design of our interface, in that we may have made it easier to click buttons corresponding to a good score. It seems however more likely due to mental fatigue. We ask our judges to look for a number of qualities—expertise, geographic appropriateness, and others—which they use to moderate their score beyond mere topicality. (Topicality, too, is already a complex notion.) This assessment does take some effort, and a tired judge may find it easier to assume a document is adequate than to check thoroughly. As evidence for this, we also see the time taken to produce a label shortens as a session goes on (−0.16 s per label, $F(1) = 2127.8, p \ll 0.001$, when modelling judge as a random variable). This is consistent with observations of TREC judges [13]. Taken together with the convergence of scores, this seems not to indicate growing fluency but decreasing mental work for each label. A similar effect was seen by Cai et al. [12] and Zhang et al. [60], who noted increasing boredom and fatigue although (in their settings) mixed impacts on result quality.

4 JUDGES HAVE THE USUAL HUMAN BIASES

A cognitive bias is a systematic form of error, resulting from applying heuristic principles when making complex decisions. For example, Tversky and Kahneman [58] discuss a bias whereby "people assess the frequency of a class or the probability of an event by the ease with which instances or occurrences can be brought to mind". This "availability bias" leads to systematic over-estimation of the probability of an event where we are familiar with some instances, regardless of whether our experience is representative. In the years since Tversky and Kahneman's work, other biases have been well documented, and their number has grown: in their 2014 review, Fleischmann et al. [21] suggest 120 identifiable biases at work in information tasks, and other scholars have listed 180 [10].

These biases have been observed in information-seeking. Behimehr and Jamali [8] give a thorough enumeration of 28 cognitive biases seen in information-seeking behaviour of graduate students, across stages from task definition to evaluation. Several biases noted there might be pertinent to crowd labelling, including the "illusion of truth" effect whereby unambiguous information is preferred; confirmation bias, a preference for information concordant with our prior beliefs; and the picture superiority effect, a preference for graphical displays over text.

Some biases have also been observed in relevance labelling in particular. Eickhoff [17] used specially-designed judge interfaces to demonstrate that a number of cognitive biases can be induced in relevance rating. By controlling the documents shown and the way they were presented, Eickhoff was able to show an ambiguity effect (whereby documents with missing metadata were rated lower); an anchoring effect (where judges who had labelled a document based on random, fake, metadata did not adjust their rating when shown the real content); a bandwagon effect (where judges agreed with others, even if those "others" were fake); and a decoy effect (where seeing a third document changed perceptions of the first two). These biased labels led to somewhat different system scores and somewhat different system-level rankings. However, these effects were shown by manipulating judging interfaces and assignments, and therefore must represent extreme cases. In our analysis we use a production system without manipulation to show similar biases.

Azzopardi [4] catalogued and reviewed other work on cognitive biases in IR. He suggests that in relevance labelling, we may see biases due to anchoring, availability, trust, ambiguity, priming, and decoy effects. Our data confirm some of these effects and we explore others too.

4.1 Anchoring

Anchoring is a process whereby, when making (apparently) absolute judgements on some dimension, humans can be swayed by an earlier reference or "anchor" in that same dimension. The effect has been demonstrated repeatedly, across many domains, and seems to be deep-seated and hard to overcome [59].

Anchoring effects have also been observed in relevance judging, although results have been inconsistent. Scholer et al. [53] controlled the quality of the first twenty documents seen by a judge, in three buckets (high, medium, and low quality), and observed a corresponding shift in labels for the next 28 documents. Labels shifted lower given the high-quality prologue, and higher given the low-quality prologue, to a net difference of around 0.3 points on a four-point scale.

Shokouhi et al. [55] also demonstrated an anchoring effect in relevance assessments, again of moderate overall impact, but in the opposite direction to Scholer et al. Judges were given pairs of documents for a single query, where the first document was of known quality. Judges given a "perfect" document first were substantially more likely to use the same label for their second document, and similarly for "good" first documents. For "bad" first documents, there was instead a skew to "good" in the second label. Carterette and Soboroff [13] reported a similar autocorrelation in TREC judgements, where labelling one document relevant increased the chance of the next document getting the same label by a relative 22%. Newell and Ruths [46] saw related "negative priming" effects in image labelling.

[2] Other breaks, from 30 minutes to several hours, make little difference to our conclusions.

Table 1: Relative abundance of labels, on a five-point scale, by the first label given in a session. With the exception of “fair”, all labels are relatively more abundant if they match the first label in a session; and with that exception, all labels show the biggest change in abundance if they match the first label in a session.

	Subsequent label				
First label	Bad	Fair	Good	Excellent	Perfect
Bad	+9.19 %	−3.43 %	−4.59 %	−2.52 %	−13.88 %
Fair	+6.84 %	−6.90 %	−0.24 %	−3.22 %	−18.09 %
Good	−1.55 %	−5.62 %	+6.84 %	−3.06 %	−6.22 %
Excellent	−1.51 %	+5.77 %	−4.62 %	+4.65 %	−8.58 %
Perfect	−12.97 %	+10.18 %	+2.61 %	+4.15 %	+46.78 %

To assess the effect anchoring might have on our relevance labels, we used a similar data set to that above with over 300,000 labels collected from almost 600 judges. Again, labels were on a five-point scale. After breaking into sessions as before, we recorded the first label given in each session—the anchor—as well as the labels coming later in the session.

Table 1 tabulates our results, for each label seen first in a session. There is a clear correlation: for example, if the first label in the session was “bad”, subsequent labels in the session were 9% more likely to be “bad” and 14% less likely to be “perfect”.

In our setup, the value assigned to the anchor is generated by the judge themselves or “self-generated” [19]. We are not communicating to say “this document is bad”, and indeed the randomisation of tasks means we have no control over the order that documents or queries are seen. As a result, we would expect only a very small effect due to anchoring. However, the effect is clearly noticeable: in almost every case, if the first document in a session is labelled X then subsequent Xs are more likely. (The exception is the “fair” label, which is used relatively infrequently and for documents of middling quality.)

In spite of the difference in setting, this is a similar observation to that of Shokouhi et al. [55], who used a three-point scale and controlled the quality of the anchor, but it differs considerably from the observations of Scholer et al. [53]. It is probably salient that our data looks different in several ways: Scholer et al. considered longer-term effects (the 21st label and beyond, given an anchor of 20 documents), within a lengthy session on a single topic; Shokouhi et al. considered immediate pairs of documents, again on a single topic; and we look at the first document in arbitrarily-long sessions, with mixed topics. We also did not control the quality of the first document seen.

Epley and Gilovich [19] distinguish “externally provided” from “self-generated” anchors. In the latter case, they suggest, we adjust from the anchor value according to any perceived differences between the objects being judged. Here, that would suggest that judges make a reasonable guess at the quality of the first document they see, then look for differences in quality given the next document and adjust their score accordingly. If those differences were difficult to detect, or difficult to decide on, or simply small—as would happen fairly often—then we would see the effect described here. Our data is therefore at least consistent with Epley and Gilovich’s model, whereas that of Scholer et al. would be consistent if the difference in quality were easier to notice.

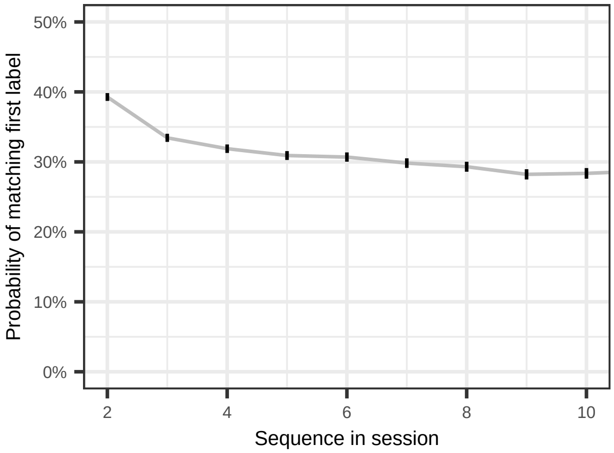


Figure 4: The probability of a label matching the first given in the judging session. The second label in a session is rather more likely to match the first, then the probability of a match decays to no more than chance. Error bars denote ±1 SE.

Further evidence for an effect. One objection is obvious. A search engine will naturally tend to produce good (or bad) results in general, meaning everything seen in a judging session is likely to be good (or bad), and labels would naturally correlate across the session (and not just with the first).

We have tried to control for this by analysing only labels which came from production systems, during particularly busy labelling periods when results of tens of thousands of queries were being labelled at the same time. By using production ranking algorithms, results will not be uniformly bad; by using more than one ranking algorithm, and tens of thousands of queries, we will see a range of performance and a range of document relevance in the judging pool. Since queries, ranking algorithms, and documents are finally randomised across judges and within a session, it seems unlikely that a judge could get documents of non-random quality across a judging session.

It is still possible, however, that some engines are consistently good or bad over some aspect which is not randomised. An obvious candidate here is language: we do not randomise over languages and of course each judge is only asked to label results in a language they can read. If an engine were good (or bad) in any one language, and documents in those languages were batched for judging, we may see misleading correlations.

Figure 4 addresses this by plotting, for each label from the second in a sequence, the chance that it is the same as the first. Since documents and queries are randomised in a session, if there were no anchoring effect then we should expect the distribution of labels to be randomised as well: that is, every label should have the same chance of agreeing with the first and this plot should be a flat line.

What we see instead is that the second label in a session is more likely to match than is the third; the third more likely than the fourth; and so on until the odds stabilise some time after the tenth label, meaning the first label given and first document seen no longer have any influence. This is a clear sign of an anchoring effect, diminishing over time as judges see more examples of documents and perhaps better calibrate their own scales. Again, this is consistent with Shokouhi et al. [55] and Carterette and Soboroff [13], although extended to longer intervals.

4.2 Mere exposure

The mere-exposure, or familiarity, effect holds that people develop a preference for a stimulus simply by repeated exposure (see Bornstein [11] for an overview). Azzopardi [4] suggests this effect might be seen in interactive retrieval: for example, well-known sites might be preferred more than they "ought" on the basis of utility alone. To our knowledge, this has not been investigated in relevance labelling.

Again, we do see evidence of this in our data: judges gave more favourable scores to web documents from hosts that they saw more often. To model the effect of exposure, each URL in our data set was reduced to the host name, and we counted how often each host name had been seen by the relevant judge at the time the web document was scored. A mixed-effects model again showed a positive effect of number of exposures on score (0.5 points per exposure, $t(529\text{k}) = 20.36, p \ll 0.01$). Each time the same host was seen in a session, it attracted a higher score, e.g., the fifth time a host was seen it was scored on average 2 points (of 100) higher than the first.

In our setup, the web documents from any one host are presented in more or less random order. Further, as hosts are seen across multiple sessions, exposure count has essentially no correlation with time in session (Pearson $r = 0.02$, $t(529\text{k}) = 13.16, p \ll 0.01$). We therefore believe that the two effects—of time and of exposure—are largely independent. There is no reason to expect that URLs from one host get better over time, and exposure or familiarity seems a likely explanation for the observed increase in scores.

4.3 Left-side bias

Test-collection-based methods have one great disadvantage: by treating the result-level relevance labels as independent, most metrics are blind to the diversity or redundancy across a retrieved set. That is, a ranking algorithm which returns near-identical documents repeatedly might score better than one which returns a greater variety. One way to counter this is to ask judges to label an entire set or even an entire search engine result page (SERP) [6]. Commonly, this is done side-by-side, where judges are shown two sets of results for the same query, adjacent to each other, and are asked which they prefer [56].

In other fields, there is evidence that, given choices laid out left to right, research participants tend to prefer the left-hand options [22]. The effect is inconsistent [41] and varies across scales and populations, but has been repeatedly observed with small magnitude [40].

In our case, we do observe a clear left-right bias. In one example, the pair of SERPs in Figure 5 was presented to 1000 crowd judges—in this case, untrained and novice—with 500 getting each ordering. The SERPs differed only in that one had the traditional white background; the other had a neon green background, which was jarring and which made the text hard to read. Despite this, a full 32% of respondents claimed to prefer the green when the green SERP was shown to the left. With the green SERP on the right, this fell to 21% ($\chi^2(1) = 14.324$, $p = 0.0002$): a relative 51% benefit from being shown on the left-hand side.

In a less dramatic example, we collected over 50,000 labels representing preferences over around 5000 pairs of SERPs; each pair was presented five times with one SERP on the left, and five times with the other SERP on the left. Judges in this case were experienced at the task: nevertheless, preference within each SERP pair correlated with which was on the left-hand side.[3] SERPs appearing on the left benefited from about a 2% increase in preference ($\chi^2(1) = 4.92$, $p = 0.03$). A 2% shift is smaller than in the previous experiment, which we attribute to the difference in judge groups. However, 2% is in the range reported by Lewis and Sauro [40], and of course we have a different setting—and larger data—than the studies surveyed there. This bias is probably too small to be interesting if the SERPs are radically different, but is certainly material when the effects being studied are already small (for example, as the result of small system changes; or ill-defined qualities, such as utility or relevance, which are already hard to agree on).

5 JUDGES SWITCH TASKS

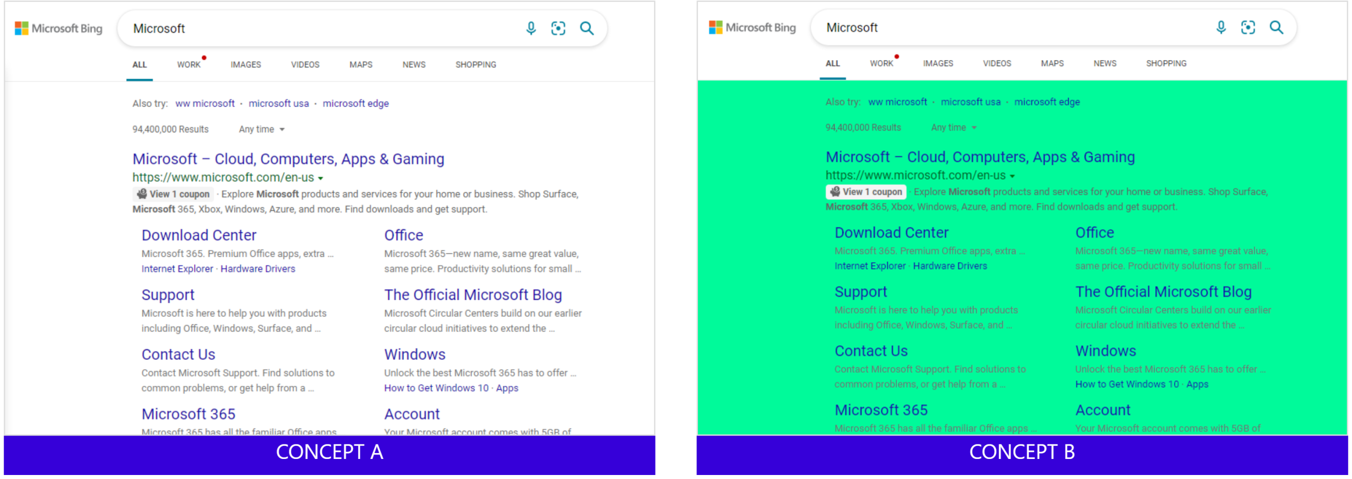
In most computer-based work there is a tendency to multitask [26, 42]. This has been observed in crowdsourcing systems as well, with workers on Amazon Mechanical Turk switching tasks every 5 minutes on average [28]. This is likely due to boredom—especially given the repetitive nature of most crowd tasks [60]—but does reduce quality overall [34, 38]. We have observed similar patterns amongst our crowd judges.

Our labels are collected on our own in-house platform (UHRS), hosting a variety of tasks beyond the relevance labelling described above: classification, transcription, and annotation of various media for example, as well as surveys and studies with interactive systems. For the analysis here, we collected well over 300 million microtasks ("HITs"), such as providing a single relevance label or classification. These were drawn from almost 3000 different "applications", each collecting different kinds of labels or operating with different interfaces, judges, languages, or regions. The HITs were completed by well over 100,000 judges.

We study judges' daily engagements with HIT applications. On average, judges engaged with 2.14 applications per day, and only 28% of judges worked on a single application each day. This could of course mean that judges worked on separate tasks hours apart; but there is ample task switching in shorter windows, and more than half our judges engage with more than one application in any twelve-minute window.

Judges seem to prefer switching between two applications, which could mean that they have adopted a strategy to reduce the delay they incur while waiting for a new microtask to load. Other reasons for switching may include if work on one application has run out; if one application is simply not to the judge's taste; or if the judge was automatically blocked after falling short of some quality threshold, i.e., if the judge's work is considered unreliable.

[3] This data was collected from English-speaking judges, with English SERPs.



Microsoft

Also try: ww microsoft · microsoft usa · microsoft edge

94,400,000 Results Any time

Microsoft – Cloud, Computers, Apps & Gaming
https://www.microsoft.com/en-us
View 1 coupon · Explore **Microsoft** products and services for your home or business. Shop Surface, **Microsoft** 365, Xbox, Windows, Azure, and more. Find downloads and get support.

Download Center
Microsoft 365. Premium Office apps, extra ...
Internet Explorer · Hardware Drivers

Office
Microsoft 365—new name, same great value, same price. Productivity solutions for small ...

Support
Microsoft is here to help you with products including Office, Windows, Surface, and ...

The Official Microsoft Blog
Microsoft Circular Centers build on our earlier circular cloud initiatives to extend the ...

Contact Us
Contact Microsoft Support. Find solutions to common problems, or get help from a ...

Windows
Unlock the best Microsoft 365 has to offer ...
How to Get Windows 10 · Apps

Microsoft 365
Microsoft 365 has all the familiar Office apps

Account
Your Microsoft account comes with 5GB of

Microsoft

Also try: ww microsoft · microsoft usa · microsoft edge

94,400,000 Results Any time

Microsoft – Cloud, Computers, Apps & Gaming
https://www.microsoft.com/en-us
View 1 coupon · Explore **Microsoft** products and services for your home or business. Shop Surface, **Microsoft** 365, Xbox, Windows, Azure, and more. Find downloads and get support.

Download Center
Microsoft 365. Premium Office apps, extra ...
Internet Explorer · Hardware Drivers

Office
Microsoft 365—new name, same great value, same price. Productivity solutions for small ...

Support
Microsoft is here to help you with products including Office, Windows, Surface, and ...

The Official Microsoft Blog
Microsoft Circular Centers build on our earlier circular cloud initiatives to extend the ...

Contact Us
Contact Microsoft Support. Find solutions to common problems, or get help from a ...

Windows
Unlock the best Microsoft 365 has to offer ...
How to Get Windows 10 · Apps

Microsoft 365
Microsoft 365 has all the familiar Office apps

Account
Your Microsoft account comes with 5GB of

Figure 5: A pair of SERPs differing only in their background colour: conventional white or hard-to-read bright green. The green was preferred substantially more often when it was shown to the left.

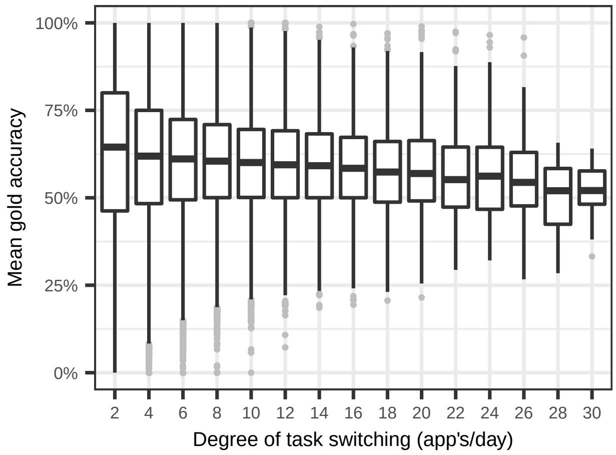


Figure 6: Quality of labels decreases as the degree of judge task switching increases. “Task switching” is measured as the number of applications a judge contributed to in a day; quality is measured by agreement with gold labels.

In common with other crowd systems, our platform allows task owners (employees and contractors) to set quality thresholds for crowd judges. If a judge’s work falls under this threshold, they may be temporarily or permanently barred from submitting more work.

Thresholds are commonly set on one or both of two criteria: first, the degree of agreement between a judge’s labels and some predetermined “gold” labels, provided by task owners; and second, the degree of agreement between the judge and their crowd peers, across those instances where more than one judge completes the same microtask [33].

On three measures—agreement with gold labels, pairwise agreement with other judges, and number of times in a day that a judge is blocked—we see a negative association between degree of task switching and quality of labels. As the degree of judge task switching increases, we see a decline in judge accuracy—measured as the proportion of gold HITs, seen by that judge that day, where they gave the required answer (Figure 6). This effect is statistically significant (a 0.3% reduction in accuracy per simultaneous application, ANOVA $F(1) = 1506.2$, $p \ll 0.001$). We see something similar with pairwise accuracy, or the proportion of those microtasks where the judge agrees with their peers (−0.06% accuracy per simultaneous application, ANOVA $F(1) = 174.78$, $p \ll 0.001$).[4]

As final evidence of a link, we see that the number of applications a judge attempts in a day correlates with the number from which they are blocked (0.2 applications blocked per application attempted, $F(1) = 904\text{k}$, $p \ll 0.001$).

Finding that workers who attempt more tasks tend to have lower-quality output agrees with Gould et al. [28]. We do not claim a *causal* effect: it may be that workers who switch tasks are inattentive, or it may be that workers tend to switch tasks when they feel they are not performing well. In the latter case, this switch may be motivated by a desire to do well, or an aversion to getting banned. Judges of course may fail a quality bar not for malicious reasons, but simply because they are trying to learn a new task and trying to understand what is expected of them. It is also possible that task owners set unreasonable expectations and too stringent quality thresholds. This can easily happen since the owners likely have a very deep understanding of their own tasks, and therefore struggle to anticipate the pitfalls and areas of confusion. From our current study, we observe the correlation but cannot distinguish these various explanations. We note however that most judges’ incentive is to get through as much work as possible, without getting flagged as a poor worker or as a spammer, and task switching is a rational decision inside the systems we build. More worker-friendly strategies may try to address poor performance by offering training or

[4] Gold HITs are stratified in varying ways across these applications and we do not make any claim about judges’ absolute accuracy overall. Our focus is the trend in gold accuracy, as task switching increases.

seeking feedback, with the aim to identify task design issues that may have been at the root cause of the poor quality output.

6 JUDGES DON'T AGREE ANYWAY

Finally, "relevance" itself is notoriously hard to define and has been the subject of analysis and (re-)definition over decades [51, 52]. Distilling this complex notion to a usable scale, while taking into account the variety of web search tasks and of web results, is extremely difficult. Google's "search quality evaluator guidelines", which judges are expected to digest, run to over 170 pages [27]; Bing's guidelines are reported to run for 70 pages.[5]

Individual judges accordingly vary considerably in their labelling practice, and this effect often dominates the effects due to task switching or anchoring, or even the larger effects due to time. For example, in our model of session length, we saw an effect of 0.013 points per judgement in a session. Allowing for this, however, still left a very large effect due to judge: per-judge intercepts ranged from 7 to 79 points on our 100-point scale, and there is a 13-point difference from the first to the third quartile. In our model of time of day, we saw a much bigger effect of up to almost six points (Figure 2), but again judge effects are large even accounting for this (a 15-point inter-quartile range).

Judges can also disagree with themselves—that is, they can be inconsistent. In their work with TREC labels, Scholer et al. [54] found that when a judge saw a document which was a near-identical copy of one they had seen earlier, for the same topic, there was a 15–24% chance they would assign a different label. (The chance varied according to collection, and hence according to the type of document and the labelling scheme.) In some, but not all, document collections this chance increased as more time passed between duplicates. Scholer et al. considered this evidence either that judges tended to forget guidelines, or that their view of relevance changed as time went on. We add that this might also be due to any of the factors listed above. Scholer et al.'s effect was big enough to change system orderings, even after aggregating over queries. We do not have comparable data in our collection, but there is no reason to believe our judges would be any more self-consistent.

7 CONCLUSIONS AND RECOMMENDATIONS

Just like many other researchers, we rely on crowd workers labelling web documents as a critical part of our evaluations. These crowd workers are subject to the same cognitive biases, shortcuts, and distractions as anyone else, and we see evidence of this in the labels we gather. These effects include large swings in labels across the day and week; fatigue during a labelling session; anchoring and mere-exposure effects; and left-side bias. We also see evidence that task switching introduces inaccuracy in labelling, although crowd systems almost inevitably drive workers to this; and evidence that simple disagreement introduces large differences between workers.

Table 2 summarises our results. The *scale* of these effects is specific to our tasks, ranking algorithms, queries, platform, scale, and labelling interface—amongst other factors—and effects will also vary more or less by specific individuals or cohorts of judges. The *existence* of these effects, however, we expect to be common across most crowd labelling, especially relevance labelling.

7.1 Interactions

In our analysis, we have examined the effects independently. These effects, of course, are very likely to interact: we might, for example, see more order effects when a judge is already tired, or there might be an interaction between time of day, day of week, and judge location (for example, around what counts as a weekend, or what are normal working hours). Any of these could further skew crowd-based evaluations.

The data we used was pulled from a range of different experiments, so robust analysis of all interactions is not possible here. There are also, of course, very many possible interactions; the resulting models quickly get very complex, and accordingly harder to interpret and use.[6] We can however summarise some examples:

Fatigue and time of day. We previously showed an effect due to time in the judging session, whereby scores tend to get higher over time (Section 3). This effect interacts with the (judge-local) hour of the day, such that the effect is largest at either end of the workday (23 times higher at 8am and 5pm, 15 times higher at 9am, 7 times higher at 6pm) and reverses in the middle of the night (−16 times at 10pm for example, through to −11 times at 4am; all specified interactions statistically significant per ANOVA F, $p < .05$, judge as a random effect). That is, scores tend to inflate as people work longer during the day, but deflate the longer people work overnight. This may be due to a combination of fatigue and underlying effects.

Time of day and day of week. We also see an effect on label due to the interaction between time of day and day of week. On weekends (Saturday and Sunday), there is less negative effect in the late night and early morning, but lower labels on average during the late afternoon and early evening (ANOVA F varies, $p < .05$).

Given the possible complexity of these interactions, further effort is needed to define a subset which are interesting prima facie: ideally, we would list interactions we expect to be substantial and also which we can control, to debias our labels. Testing for and modelling these interactions, and testing the debiasing, we leave for future work.

7.2 Recommendations

From our observations above, we can propose a small set of principles for crowd relevance labelling. Evaluation is almost always a comparison—comparing two ranking algorithms, or two variants of one ranking algorithm—so we want the labels on which our scores are based to vary only according to the quality of the ranking algorithms' results. This means we must minimise other effects, including those discussed above. Draws et al. [16] offer a useful twelve-item checklist of cognitive biases to watch for, and suggest researchers (1) assess, (2) mitigate, and (3) document these. To that list, we would add further recommendations to address the biases and observations above:

[5] See for example https://blog.searchevaluator.com/web-content-assessor-lionbridge/, reporting "about 70 pages" in August 2020.

[6] For example, modelling the interaction between day of week, hour of day, and judge location leads to well over 3000 effects.

Table 2: Summary of the effects discussed in our examples. The data here is specific to our tasks, judges, ranking algorithms, queries, etc; but the biases represented here are likely to be observed in any large-scale crowd labelling programme.

Factor	Effect
Document judging (100-point scale)	
Time of day	−3.79 to +2.71 points
Day of week	−1.02 to +1.22 points (but varies by region)
Time in session	+1.09 per hour (and variance reduces)
Sequence in session	+0.013 per score given (and variance reduces)
Exposure to host	+0.5 per exposure
Anchoring on judge's first label	varies across a session
SERP-pair preference judging	
Left-side bias	2%–51%
All crowd tasks	
Task switching	−0.3% agreement per application in a day

(1) Minimise the effect of time and day by releasing labelling jobs at the same time whenever possible, and especially by releasing jobs for all ranking algorithms simultaneously; *and*
(2) Include some overlap between batches (some documents judged at more than one time), to observe the difference.
(3) In side-by-side comparisons, always balance the two views: that is, have each system be both shown in equal proportions on the left and on the right over time.
(4) Make changes to the judgement process to address anchoring, fatigue, and task switching effects, where possible. The first can be minimised by mixing queries in the same session, so documents are clearly different and judges come closer to making absolute judgements. To control the second, it may be worthwhile capping the number of documents labelled in one session or in one day. Task switches are hard to control unless we control the platform, and will be only partially controlled even then, and is a crude response to designing slow and/or boring tasks in the first place. Happily, observed effects due to task switching are small.
(5) Measure and be aware of per-judge differences; consider normalising for this.
(6) Importantly: regardless of what mechanism(s) are used, look for, measure, and report these sorts of effects and measure and report the efficacy of any controls.

7.3 Conclusions

We conclude with three observations.

Effect on production metrics. First, the various effects above—especially the large effects due to judge priors, time of day, and day of week—have led to material changes in production metrics, at scale in Bing. The effect which we attribute to the religious festival may be obvious from looking at the data; however, the root cause took some time to identify (and our hypothesis is impossible to prove), and in the interim our metric was unreliable. Effects due to time were not as obvious day to day, although from our experience we knew to look for these after seeing errors in unrelated, pre-production, systems. The other effects are more subtle and we may never have noticed them without careful checking. Regardless, these all influence large-scale measurement.

Effect on research evaluations. The exact shape and size of these effects is of course particular to our tasks and our crowd, but we expect similar effects elsewhere—especially since these are manifestations of broad human biases, shortcuts, and physiological factors. That suggests that metrics elsewhere, including in the research literature, may be impacted. Unlike production metrics, there is not likely to be experience from running the same evaluations repeatedly. We could quite easily miss any bias, until and unless we routinely test for these effects and report appropriately. This is important for large-scale exercises, where labelling may involve many people or may run for days at a time. It is especially important where this may interact with an imbalance between systems under test—for example, if one system is tested later than another, so day-of-week effects shift the two sets of labels differently. This is common if one system did not contribute to the original pool, or if the pool is not randomised before labelling.

Emphasis and terminology. Finally, we observe that it is common, on seeing crowd data deviating from our expectation, to refer to "noise" or "judge error" [for example 13, 36, 54]. Some of the deviation may fairly be so called; but there is a reasonable fraction which has some structure, which is therefore predictable to some extent, and which can be explained by the usual human biases and incentives. It is not an "error", nor simply "noise", to be subject to anchoring, exposure, or to get tired or hungry. We need to account for the human processes involved in labelling, and the predictable ways these might influence our results. At the end of the day, "the crowd" is comprised of people.

7.4 Future work

Our work has shown that there are many human factors that can impact the use of crowd labelling for search evaluation and affect the results that are obtained. Future work includes expanding the scope of this work to include other effects (e.g., additional cognitive biases [4] and other physiological factors such as sleep quality, which can be estimated at scale in search settings [3] and can affect cognitive performance). It should also include interactions between effects, as well as tasks in the IR domain beyond search result evaluation, e.g., query classification. We also need to perform additional analysis, including using causal inference methods [31] to more fully understand causality in settings such as task switching or multitasking, and additional experimentation to focus on specific research questions of interest, e.g., variations in judging per individuals or cohorts or effects over multiple judging sessions. Overall, it is clear that better understanding factors that affect the crowd is a rich and important research area that needs more investigation. We hope that this initial foray paves the way for considerable additional study.

ACKNOWLEDGMENTS

Shiroy Choksey, Deepak Santhanam, and Trudy O'Connor offered insights and data. The anonymous reviewers provided useful suggestions for improving this work. We remain grateful for the time and effort of the thousands of real people in our "crowds".

REFERENCES

[1] Omar Alonso and Ricardo Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *Proceedings of the European Conference on Information Retrieval*. 153–164.

[2] Omar Alonso, Daniel E Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. In *ACM SIGIR Forum*, Vol. 42. 9–15.

[3] Tim Althoff, Eric Horvitz, Ryen W White, and Jamie Zeitzer. 2017. Harnessing the web for population-scale physiological sensing: A case study of sleep and performance. In *Proceedings of the 26th International Conference on the World Wide Web*. 113–122.

[4] Leif Azzopardi. 2021. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*. 27–37.

[5] 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 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 667–674.

[6] Peter Bailey, Nick Craswell, Ryen W White, Liwei Chen, Ashwin Satyanarayana, and Saied MM Tahaghoghi. 2010. Evaluating search systems using result page context. In *Proceedings of the 3rd Information Interaction in Context Symposium*. 105–114.

[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.

[8] Sara Behimehr and Hamid R. Jamali. 2020. Cognitive biases and their effects on information behaviour of graduate students in their research projects. *Journal of Information Science Theory and Practice* 8, 2 (2020), 18–31.

[9] Erik M Benau, Natalia C Orloff, E Amy Janke, Lucy Serpell, and C Alix Timko. 2014. A systematic review of the effects of experimental fasting on cognition. *Appetite* 77 (2014), 52–61.

[10] Buster Benson. 2016. Cognitive bias cheat sheet. https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18. Downloaded: 2022-01-08.

[11] Robert F Bornstein. 1989. Exposure and affect: overview and meta-analysis of research, 1968–1987. *Psychological Bulletin* 106, 2 (1989), 265.

[12] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 3143–3154.

[13] Ben Carterette and Ian Soboroff. 2010. The effect of assessor errors on IR system evaluation. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 539–546.

[14] Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2013. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing* 17, 4 (2013).

[15] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 135–143.

[16] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 48–59.

[17] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 162–170.

[18] Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 2012. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 871–880.

[19] Nicholas Epley and Thomas Gilovich. 2005. When effortful thinking influences judgmental anchoring differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making* 18, 3 (2005), 199–212.

[20] Ozkan Eren and Naci Mocan. 2018. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics* 10, 3 (2018), 171–205.

[21] Marvin Fleischmann, Miglena Amirpur, Alexander Benlian, and Thomas Hess. 2014. Cognitive biases in information systems research: A scientometric analysis. In *Proceedings of the European Conference on Information Systems*.

[22] Hershey H Friedman and Taiwo Amoo. 1999. Rating the rating scales. *Journal of Marketing Management* 9, 3 (1999), 114–123.

[23] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Learning Analytics and Knowledge Conference*. 105–114.

[24] Alyssa A. Gamaldo, Jason C. Allaire, and Keith E. Whitfield. 2010. Exploring the within-person coupling of sleep and cognition in older African Americans. *Psychology and Aging* 25, 4 (2010), 851–857.

[25] Scott A. Golder and Michael W. Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333, 6051 (2011), 1878–1881.

[26] Victor M. González and Gloria Mark. 2004. Constant, constant, multi-tasking craziness. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 113–120.

[27] Google. 2020. Search quality evaluator guidelines. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf. Downloaded 2021-10-11.

[28] Sandy J. J. Gould, Anna L. Cox, and Duncan P. Brumby. 2016. Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction* 23, 3, Article 19 (June 2016).

[29] Javier Hernandez, Mohammed (Ehsan) Hoque, Will Drevo, and Rosalind W. Picard. 2012. Mood meter: counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 301–310.

[30] Danula Hettiachchi, Niels van Berkel, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2019. Effect of Cognitive Abilities on Crowdsourcing Task Performance. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. 442–464.

[31] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81, 396 (1986), 945–960.

[32] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1–12.

[33] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the SIGKDD Workshop on Human Computation*. 64–67.

[34] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P. Bigham. 2018. Striving to earn more: A survey of work strategies and tool use among crowd workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 70–78.

[35] Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Proceedings of the European Conference on Information Retrieval*. 165–176.

[36] 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 21st ACM International Conference on Information and Knowledge Management*. 105–114.

[37] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82, 13 (2017), 1–26.

[38] Laura Lascau, Sandy J. J. Gould, Anna L. Cox, Elizaveta Karmannaya, and Duncan P. Brumby. 2019. Monotasking or multitasking: Designing for crowdworkers' preferences. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1–14.

[39] Matt Lease and Emine Yilmaz. 2013. Crowdsourcing for information retrieval: Introduction to the special issue. *Information Retrieval Journal* 16 (2013), 91–100.

[40] Jim Lewis and Jeff Sauro. 2020. Revisiting the evidence for the left-side bias in rating scales. https://measuringu.com/revisiting-the-left-side-bias/. Downloaded 2021-09-09.

[41] James R Lewis. 2019. Comparison of four TAM item formats: Effect of response option labels and order. *Journal of Usability Studies* 14, 4 (2019), 224–236.

[42] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1739–1744.

[43] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a Turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. 224–235.

[44] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V. Hanrahan. 2016. Turking in a Global Labour Market. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing*. 39–77.

[45] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1 (Dec. 2008).

[46] Edward Newell and Derek Ruths. 2016. How one microtask affects another. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 3155–3166.

[47] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Just the right mood for HIT! Analyzing the role of worker moods in conversational microtask crowdsourcing. In *Proceedings of the International Conference on Web Engineering*. 381–396.

[48] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[49] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 2863–2872.

[50] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 248–375.

[51] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 1915–1933.

[52] Tefko Saracevic. 2012. Research on relevance in information science: A historical perspective. In *Proceedings of the ASIS&T Pre-conference on the History of ASIS&T and Information Science and Technology*. 49–60.

[53] Falk Scholer, Diane Kelly, Wan Ching Wu, Hanseul S. Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 623–632.

[54] Falk Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1063–1072.

[55] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 963–966.

[56] Paul Thomas and David Hawking. 2006. Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. 94–101.

[57] Ming-Chang Tsai. 2019. The good, the bad, and the ordinary: The day-of-the-week effect on mood across the globe. *Journal of Happiness Studies* 20, 7 (2019), 2101–2124.

[58] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185 (1974), 1124–1131.

[59] Timothy D. Wilson, Christopher E. Houston, Kathryn M. Etling, and Nancy Brekke. 1996. A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General* 125, 4 (1996), 387–402.

[60] Ying Zhang, Xianhua Ding, and Ning Gu. 2018. Understanding fatigue and its impact in crowdsourcing. In *Proceedings of the 22nd IEEE International Conference on Computer Supported Cooperative Work in Design*. 57–62.