We thank the reviewers for their constructive feedback and suggestions for improvements. Our responses to particular comments are outlined below. Changes in the manuscript have been highlighted in red font to make them easy to spot.

1 Reviewer 1

In the footer on the first page, please add a concise statement of the paper's addition after "This paper is a revised and extended version of [Turpin et al. 2015]; it adds XXX." By the way, you should use [Author year] when you refer to a publication and Author [year] when you refer to the author (citepxxx and citetxxx, respectively).

The citation style has been fixed, thank you. A detailed list of changes has been added at the end of Section 2.2, in line with the next item.

In the related work section, please make precise how the TOIS submission differs from the earlier SIGIR 2015 paper.

A paragraph has been added at the end of Section 2.2 to do this.

RQ2 is an important additional ingredient of this extended paper. Section 6 is devoted to answering this question. Please explicitly announce that that is what you are going to do. And please provide answers to the question at the end of section 6. We are now left to discover for ourselves what (if any) answer you were able to produce for RQ2.

We have added Section 6.4 which summarises the section.

2 Reviewer 2

I like the paper and the problem that it addresses: magnitude estimation for information retrieval and the use of crowd sourcing for this purpose. The authors claim that their findings are direct implications for IR evaluation. This is probably true, but I think the authors can do a better job of articulating them. I recommend that each result section 5-8 explicitly (1) starts with one of the four research questions, (2) answers it, (3) explains what the ramifications of the answers are for IR evaluation.

This has been done for Section 6 (RQ2) by adding Section 6.4, Section 7 (RQ3) by adding Section 7.4, and Section 8 (RQ4) by adding Section 8.1. We feel RQ1 is reasonably covered in Section 5.

Some more specific comments:
Section 2.3: good addition. You could consider adding a brief discussion on the value of crowd worker disagreement. There is work by Lora Aroyo and Chris Welty that you should probably relate to in section 2.3

Thanks for pointing this out, we have added a paragraph in Section 2.3.

Section 4: This is a better way of organizing the material than in the SIGIR version. Collect and report on the data first. And only interpret and conclude later. Moving bits from section 3 here (on crowd judging and score normalization) makes good sense. Figures 1 and 2 are new. And so is section 4.2; adding a sentence to 4.2 with the upshot of the section would be useful. The expansion of 4.3 (formerly 3.3) is useful.

We have added the sentence at the end of 4.2

The new section 5 coincides more or less with section 4 in the SIGIR paper. Is "We therefore next investigate judge agreement" still a useful way of ending this section, given the inclusion of a new research question?

We have deleted the offending sentence at the end of Section 5.

Section 6: ranges of scores for each worker involved was very wide → RANGE of scores.

Fixed, thanks.

Figure 7 has no legends (neither x nor y).

Curious: there are axis labels, and the symbols are explained in the caption. Perhaps a printing error on the reviewer's side?

What are the practical implications of the statement at the very end of section 6, about the number of required workers? And what does it say about setups where a smaller number of workers has been used?

We have added Section 6.4 to draw this section together.

Section 7 coincides with the old section 5. Figs 10, 11, 12 need legends.

Again, this seems fine on our copies.

It would be good to explain the differences in absolute numbers between Table IV in the TOIS submission and Table 2 in the SIGIR paper. And why these differences do not matter.

We have added two sentences at the end of the first paragraph in Section 7.2 explaining the difference. On closer inspection, we also noticed that the new methodology did make a difference, and have altered the text in the final two paragraphs of Section 7.3 accordingly.

Section 8: I am not sure I am happy with the uncertain ending of this section. What does the disagreement with Kanoulas and Aslam's earlier work mean for our "ideal" evaluation setup?

We have added Section 8.1 Summary to clarify this.

Section 9: the addition of a "limitations" section is an excellent idea.
Appendix A: I like this addition. A screen dump of the instruction interface would have been helpful.

The exact instructions that appeared on the screen are already given in the appendix, and a screen dump would just format them in an HTML/www style, rather than text. As such, we don't think that an extra screen dump of the instructions would be informative. (In any case, it unfortunately turns out that CrowdFlower has undergone an 'upgrade' since we ran the experiments, so, sadly, the exact screen shot is no longer available anyway.)