



Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Palotti, Joao, [Zuccon, Guido](#), Bernhardt, Johannes, Hanbury, Allan, & Goeuriot, Lorraine
(2016)

Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions. In
International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF 2016), September 5-8, 2016, Évora, Portugal.

This file was downloaded from: <https://eprints.qut.edu.au/98863/>

© Copyright 2016 Springer

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

https://doi.org/10.1007/978-3-319-44564-9_4

Assessors Agreement: A Case Study across Assessor Type, Payment Levels, Query Variations and Relevance Dimensions

Joao Palotti¹, Guido Zuccon², Johannes Bernhardt³, Allan Hanbury¹, and Lorraine Goeuriot⁴

¹ Vienna University of Technology, Vienna, Austria,
[palotti,hanbury]@ifs.tuwien.ac.at

² Queensland University of Technology, Brisbane, Australia,
g.zuccon@qut.edu.au

³ Medical University of Graz, Graz, Austria
johannes.bernhardt@medunigraz.at

⁴ Université Grenoble Alpes, France
lorraine.goeuriot@imag.fr

Abstract. Relevance assessments are the cornerstone of Information Retrieval evaluation. Yet, there is only limited understanding of how assessment disagreement influences the reliability of the evaluation in terms of systems rankings. In this paper we examine the role of assessor type (expert vs. layperson), payment levels (paid vs. unpaid), query variations and relevance dimensions (topicality and understandability) and their influence on system evaluation in the presence of disagreements across assessments obtained in the different settings. The analysis is carried out in the context of the CLEF 2015 eHealth Task 2 collection and shows that disagreements between assessors belonging to the same group have little impact on evaluation. It also shows, however, that assessment disagreement found across settings has major impact on evaluation when topical relevance is considered, while it has no impact when understandability assessments are considered.

Keywords: Evaluation, Assessments, Assessors agreement

1 Introduction

Traditional Information Retrieval (IR) evaluation relies on the Cranfield paradigm where a test collection is created including documents, queries, and, critically, relevance assessments [12]. Systems are then tested and compared on such test collections, for which evaluation measures are computed using the relevance assessments provided. This paradigm crucially relies on relevance assessments provided by judges or annotators.

Since the inception of the Cranfield paradigm and TREC, many other evaluation initiatives have emerged (e.g., CLEF and NTCIR) and many test collections have been created. Although assessments are of paramount importance within

this evaluation method, they are often not evaluated for their reliability and are not deeply analysed. The contribution of this paper is to shed some light on this overlooked issue. With this aim, we investigate in depth the assessments of one such test collection, the CLEF eHealth 2015 Task 2 collection [9]. This collection comprises of web pages and queries issued by laypeople to find information about certain health topics, primarily with the aim of self-diagnosis. This collection fully supports the investigation of the topic of this paper because: (i) it contains two types of assessments (topical relevance assessments and understandability assessments), (ii) it contains up to three query variations for each single topic, and (iii) assessments were collected so as to have pair-wise assessments from multiple people for a set number of queries. We further add to these resources additional assessments performed by unpaid medical students and laypeople, allowing us to analyse the reliability of assessments both in terms of payment associated to the assessment task and in terms of expertise.

While some previous work has shown that large disagreement between relevance assessments does not lead to differences in system rankings [7,11], other work has shown that system ranking stability is compromised in the presence of significant disagreement [4], or large variation in topic expertise [2]. In this paper we extend prior work by considering also assessments beyond those for topical relevance and assessments with respect to query variations. In particular, because of the presence of query variations, a large proportion of documents has been judged multiple times, both within and across assessors.

2 Related Work

Prior work has examined the agreement between judges for topical relevance assessment tasks. Lesk and Salton’s work [7] is one of the earliest works on variations in relevance judgments. They found a low agreement among assessors: 31% and 33% in binary relevance assessment using Jaccard similarity to measure agreement. However, they also found that choosing one assessment or the other had little impact on systems ranking and thus rankings produced with one assessment were highly correlated to those produced with the alternative assessment. Their investigation put forward some hypotheses and reasons to justify the fact that the differences in relevance assessments did not lead to changes in system ordering. Among these were the fact that evaluation occurs over many queries and that disagreements involved mainly borderline documents.

Similar findings were reported by Voorhees [11], who studied differences in the assessments made for TREC-4 and TREC-6. For TREC-4, she used secondary assessors from NIST, while for TREC-6, she compared the assessments made by NIST with the ones made by the University of Waterloo. In both cases, the same trend unveiled by Lesk and Salton’s work [7] was found: although the agreement among judges was weak, system ordering was stable, with Kendall correlations between rankings produced using relevance assessments from different assessors varying from 89% to 95%.

Bailey et al. [2] studied three sets of assessors: “gold standard” judges, who are topic originators and experts in the task, “silver standard”, who are experts

but did not create the topics, and “bronze standard”, which are neither experts nor topic creators. They evaluated agreement among different assessment sets using conditional probability distributions and Cohen’s k coefficient on 33 of the 50 topics from the TREC 2007 Enterprise Track. Similar to the studies above, they reported little agreement between judges and, at the same time, little difference in system ordering when gold or silver judgements were used ($\tau = 0.96$ and $\tau = 0.94$ for infAP and infNDCG, respectively). However, larger differences across system rankings were observed if gold and bronze standard judgements were used ($\tau = 0.73$ and $\tau = 0.66$). This prior work supports the use of test collections as a reliable instrument for comparative retrieval experiments.

Other work has examined the impact systematic assessment errors have on retrieval system evaluation. Carterette and Soboroff [4] modified the assessments of the TREC Million Query Track to inject significant and systematic errors within the assessments and found that assessor errors can have a large effect on system rankings.





In this paper we focus on the domain-specific task of finding health information from the Web. The assessment of medical information has been shown to be cognitively taxing [6] and, as we hypothesise below, this may be one reason for disagreement on relevance assessment between and across assessors.

3 Data

In this paper we use the CLEF 2015 eHealth Task 2 dataset [9]. The dataset comprises of a document collection, topics including query variations, and the corresponding assessments, including both topical relevance and understandability assessments. Documents were obtained through a crawl of approximately 1 million health web pages on the Web; these were likely targeted at both the general public and healthcare professionals. Queries aimed to simulate the situation of health consumers seeking information to understand symptoms or conditions they may be affected by. This was achieved by using imaginary or video stimuli that referred to 23 symptoms or conditions as prompts for the query creators (see [9,10,15] for more details on the query creation method). A cohort of 12 query creators was used and each query creator was given 10 conditions for which they were asked to generate up to 3 queries per condition (thus each condition/image pair was presented to more than one person). The task collected a total of 266 possible unique queries; of these, 66 queries (21 conditions with 3 queries, 1 condition with 2 queries, and 1 condition with 1 query) were selected to be used as part of the CLEF 2015 task. A pivot query was randomly selected for each condition, and the variations most and least similar to the pivot were also selected. Examples of queries, query variations and imaginary material used for the query creation are provided in Table 1.

The collection has graded relevance assessments on a three point scale: 0, “Not Relevant”; 1, “Somewhat Relevant”; 2, “Highly Relevant”. These assessments were used to compute topical relevance based evaluation measures, such as precision at 10 (P@10), MAP and RBP. In addition, the collection also contains understandability judgements, which have been used in the evaluation to

Table 1: Example of queries from the CLEF 2015 eHealth Task 2.

Image	Information Need	Query Type	QueryId	Query Variation
	Ringworm	Pivot	03	dry red and scaly feet in children
		Most	38	scaly red itchy feet in children
		Least	45	dry feel with irritation
	Scabies	Pivot	04	itchy lumps skin
		Most	43	itchy raised bumps skin
		Least	21	common itchy skin rashes
	Onycholysis	Pivot	61	funginail bruises
		Most	19	bruised thumb nail
		Least	44	nail getting dark
	Rocky Mountain Spotted Fever	Pivot	27	return from overseas with mean spots on legs
		Most	01	many red marks on legs after traveling from us
		Least	58	39 degree and chicken pox

inform understandability-biased measures such as uRBP⁵ [14,13]. These assessments were collected by asking assessors whether they believed a patient would understand the retrieved document. Assessments were provided on a four point scale: 0, “It is very technical and difficult to read and understand”; 1, “It is somewhat technical and difficult to read and understand”; 2, “It is somewhat easy to read and understand”; 3, “It is very easy to read and understand”.

All assessments in the CLEF collection were provided by paid medical students (paid at a rate of 20 Euros per hour). We further extend these assessments by undertaking a large re-assessment exercise using a pool of unpaid medical students and a pool of unpaid laypeople volunteers. Unpaid medical students were recruited through an in-class exercise that required them to assess documents for relevance. Laypeople were recruited in our research labs: although these participants have prior Information Retrieval knowledge, they do not have any specific medical training background. The collection of these additional sets of assessments allows us to study the impact of both payment levels and expertise levels (assessor type) on the reliability of the relevance assessment exercise and system evaluation. Within this analysis, assessments performed by the paid medical students are assumed to be the gold standard. Specifically, the following relevance assessment sets (qrels) are considered in our analysis:

Default: The original set of judgements from the CLEF 2015 collection. On average, 132 documents were judged per query. Assessments were provided by 5 paid medical students.

ICS (In Class Students): The set of assessments made by unpaid medical students as an in-class activity. This set has partial assessments for 44 queries, with on average 98 documents judged per query.

Default44: The subset of documents present in the ICS set, but with assessments extracted from the Default set. This set has therefore a complete align-

⁵ uRBP is a variation of RBP [8] where gains depend both on the topical relevance label and the understandability label of a document. For more details, see [13]. In the empirical analysis of this paper, we set the persistence parameter ρ of all RBP based measures to 0.8 following [9,13].

Table 2: Comparing assessment means. Pairs that are significantly different ($p < 0.05$ using two-tailed t-test) are indicated with a star (*)

Comp.	#Top.	Asse.	Relevance	Understability	Comp.	#Top.	Asse.	Relevance	Understability
1-2	4	1	0.38 ± 0.69	$2.36 \pm 1.02^*$	ICS-1	12	ICS	$0.56 \pm 0.75^*$	$2.09 \pm 0.90^*$
		2	0.33 ± 0.64	$1.20 \pm 0.88^*$			1	$0.17 \pm 0.45^*$	$2.33 \pm 1.06^*$
2-3	3	2	$0.01 \pm 0.12^*$	$1.45 \pm 0.74^*$	ICS-2	7	ICS	$0.50 \pm 0.67^*$	$1.82 \pm 0.94^*$
		3	$0.27 \pm 0.46^*$	$2.47 \pm 0.87^*$			2	$0.02 \pm 0.15^*$	$1.16 \pm 0.87^*$
3-4	3	3	$0.62 \pm 0.62^*$	2.36 ± 0.68	ICS-3	9	ICS	$0.52 \pm 0.67^*$	$1.87 \pm 0.94^*$
		4	$0.41 \pm 0.72^*$	2.33 ± 0.72			3	$0.38 \pm 0.53^*$	$2.21 \pm 0.99^*$
4-5	3	4	$0.07 \pm 0.25^*$	$1.87 \pm 1.07^*$	ICS-4	13	ICS	$0.58 \pm 0.72^*$	1.98 ± 0.94
		5	$0.18 \pm 0.38^*$	$1.63 \pm 1.00^*$			4	$0.21 \pm 0.55^*$	1.99 ± 1.01
					ICS-5	12	ICS	$0.49 \pm 0.71^*$	$1.98 \pm 1.06^*$
							5	$0.22 \pm 0.46^*$	$1.69 \pm 0.99^*$

ment between Default and ICS and thus allows a direct comparison between paid and unpaid medical students judgements.

Laypeople: All documents of Default44 set, but judged by laypeople with respect to their topical relevance and understandability.

In the analysis reported below, we consider only the first three runs submitted by each participating team to CLEF eHealth 2015 (for a total of 42 runs), as these runs were fully assessed up to rank cutoff 10 [9].

4 Agreements for Topical Relevance Assessments

Next we analyse the agreement between assessors with respect to topical relevance and what impact this has for system evaluation. In Section 5 we shall repeat the analysis but considering understandability assessments instead.

Section 4.1 studies inter-assessor agreement across the paid medical students using a limited number of queries for which two assessors from this group both provided judgements. Section 4.2 compares the original assessments (Default and Default44) with the assessments made by unpaid medical students as in-class activity (ICS) and the Laypeople set. Section 4.3 considers the query variations included in this collection and their implications for system evaluation.

4.1 Inter-Assessor Agreement among Paid Assessors

Thirteen randomly selected queries were assigned to two assessors from the paid medical student group: 4 queries were assigned to both assessors 1 and 2, 3 queries to assessors 2 and 3, 3 queries to assessors 3 and 4, and finally 3 queries to assessors 4 and 5.

The official CLEF eHealth 2015 qrels (Default) comprises of all assessments done for queries that were judged by one assessor only; for the thirteen queries with two assessments per document, relevance labels were assigned by selecting the labels from one assessor for half of the overlapping queries, and the labels from the other assessor for the remaining half of the overlapping queries.

The left part of Table 2 shows the mean and standard deviation for assessments made by each pair of assessors for queries assessed by multiple assessors.

Table 3: Kendall’s τ correlation between systems rankings when multiple assessments are compared.

Section	Comparison	P@10	MAP	RBP
Section 4.1	Default - Inverted	0.90	0.95	0.92
	Default - Max_Label	0.94	0.93	0.95
	Default - Min_Label	0.93	0.96	0.94
Section 4.2	Default44 - ICS	0.81	0.64	0.68
	Default44 - Laypeople	0.67	0.75	0.68
	Default44 - Random	0.42 ± 0.08	0.60 ± 0.01	0.32 ± 0.09
Section 4.3	Default - Pivot	0.79	0.82	0.75
	Default - Most	0.60	0.82	0.60
	Default - Least	0.75	0.80	0.75

The means were calculated summing over all labels assigned to each document-query pair (e.g., label 2 if the document was highly relevant) and dividing the total by the number of documents in each set. Pairs that are significantly different ($p < 0.05$ using two-tailed t-test) are indicated with a star (*). From Table 2, assessors 2 and 3 exhibit a large mean difference in their assessments. This difference could be explained by the fact that the topics in common between the two assessors had very few highly relevant documents and, while Assessor 2 did not consider documents reporting differential diagnosis as “somewhat relevant”, Assessor 3 did.

Most of the pairwise comparisons in Table 2 are significantly different: how does system evaluation change if the assessments of one assessors are used in place of those of another? That is, how reliable is the evaluation (for this test collection) with respect to assessor disagreement? We study three ways to combine assessments made by the paid medical students:

1. **Inverted**: we invert the labels for the assessments chosen when two assessments were available, e.g., by assigning the label given by the other assessor (see the beginning of this section);
2. **Max_Label**: we keep the highest relevance label for any query-document that was judged by two judges;
3. **Min_Label**: similar to Max_Label, but here we keep the lowest label for any assessment made by two judges.

Table 3 reports the Kendall’s τ correlation for each of the three sets, compared to the default qrels used in CLEF eHealth 2015. The empirical results using judgements from paid medical students confirm the findings of previous studies [7,11]: assessors disagreement has little effect on system rankings and thus on their evaluation.

4.2 Influence of Assessor Type and Payment Level

In this section we compare the influence of assessor type (medical expert and laypeople) and payment level (unpaid and paid medical students). The use of unpaid assessors and laypeople has the advantage of reducing the costs associated with building the test collection, however it may come at the expense of less reliable assessments and thus system evaluation. Next we aim to determine if this issue is actually present and, if it is, how to quantify the possible error.

Unpaid and laypeople assessors used the same system used in CLEF eHealth 2015 to collect relevance assessments (Relevation [5]) and the same information displayed to paid assessors was displayed to the other assessors. However unpaid medical students had no training to use the interface (although note that the interface is intuitive) and were subjected to strict time constraints as assessments were done as an in-class activity. Laypeople had training and no time constraints.

The right part of Table 2 reports the results of the comparison between assessments in the Default set (paid medical students) and those in the ICS set (unpaid medical students). We observe that, unlike paid assessors, unpaid assessors had a strong bias towards judging documents based on their relevance to the query, rather than their relevance to the case description, which goes beyond the query and requires assessors to evaluate whether the document supports the correct diagnosis, rather than just relating to the aspects mentioned in the queries. Comparison between paid assessors and laypeople are omitted due to space constraints and are available as an online appendix; they show a similar trend to those for unpaid students.

Table 3 reports the correlation of system rankings across different evaluation measures between the Default44 assessments and: (1) ICS, (2) Laypeople, and (3) the mean correlation of 1,000 random assignment of relevance labels for all pairs of documents and queries (this represents a lower bound for disagreements and evaluation errors).

Comparing Default44 and ICS, we observe that a strong correlation (> 0.8) is found only when P@10 is used, while correlations are weaker when other evaluation measures are considered. This suggests that ICS assessments are not adequate to replace Default assessments. That is: unpaid assessors largely disagree with paid assessors with respect to relevance labels and, unlike when considering paid inter-assessor disagreement, these differences have a noticeable impact on system ranking and evaluation. This result is in line with those reported by Bailey et al. [2] when comparing gold standard assessments with the bronze standard assessments collected through crowdsourcing. We hypothesise that in our case, the consistent assessor disagreements between the two groups are due to the lack of training of the unpaid cohort for the relevance assessment task (rather than interface); a task that, for the medical domain, is rather complex [6]. Note that similar findings are observed when comparing Default44 and Laypeople assessments, with correlations between these two groups being even lower than when ICS was used (although higher than when using Random). This result further stresses the complexity of the medical assessment task and that relying on laypeople to individuate relevant documents to health-related queries can bias system evaluation, rendering it unreliable.

4.3 Assessor Agreement across Query Variations

Next we study the overlap between the assessments made for a document but with respect to different query formulations (called query variations [9,3]) collected for the same information need (case description).

First, however, we examine the distribution of documents across types of query variation (Figure 1): pivot queries, most related query (most), and least

related query (least) (see [9] for details). Query variations largely contributed new documents to the pool: every query variation was responsible for roughly a one-fold increase in the number of documents in the pool. This finding resonates with what is reported in [3].

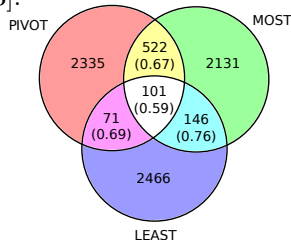


Fig. 1: Distribution of assessments for the three query variations and the agreement across pairwise types of variation.

To further quantify the role that query variations had on system evaluation, we contrast the values of mean P@10 and MAP obtained by the submitted systems on the whole set of queries, with the corresponding values obtained if only one type of query variation was used instead. Results of this analysis are reported in Figure 2 and system ranking correlations between the different settings are shown in Table 3. The highest correlation (0.82) is measured when MAP and Pivot or Most are used, while the lowest (0.60) is measured when P@10 or RBP and Most are used. For example, the plot in Figure 2 for P@10 shows that if only the Pivot variations are used, the KISTL3 run would be ranked as 2nd best, while, when all variations are used, this run is only ranked 8th. Similarly, when only the Least variations are used, KISTL3 is ranked 20th. These results suggest that using only one type of query variation does lead to noticeable different system rankings and thus the use of multiple query variations is an important aspect for system evaluation, as it more realistically captures the use of search systems than considering one type of query variation only.

There are many ways to experiment with assessments derived from query variations. Given all queries for one type of query variation, we first measure system effectiveness using the assessments for this query variation and compare with those for other variations. In Table 4, we examine whether qrels for one type of query variation can be used to assess another type of variation, e.g., use qrels for Pivot to evaluate document rankings created in response to queries from the Most variations. Given the limited document intersection between different types of queries (see Figure 1), it is expected that the correlations across different variations are small. Similar to Section 4.1, we evaluate the Min_Label and Max_Label; however, now the min and max functions are applied to the three types of query variation. Due to the larger coverage of Min_Label and Max_Label, correlations are high in most of cases.

5 Agreements for Understandability Assessments

In this section we analyse the agreement between assessors with respect to assessments of the understandability of information contained within documents and its impact on system evaluation. Understandability assessments are used

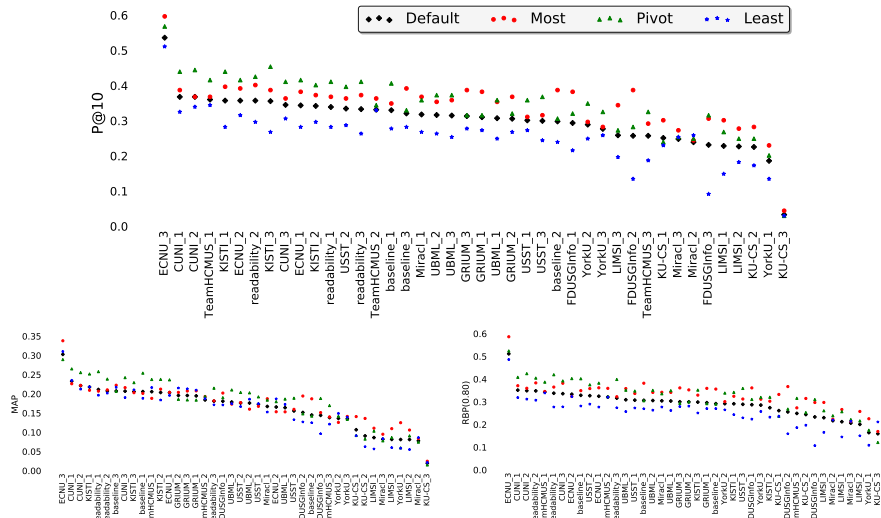


Fig. 2: System performance using queries and assessments for only one single query variant.

Table 4: Kendall- τ rank correlations for comparison of system ranking when different qrels are used.

Run Set	Qrel Comparison	P@10	MAP	RBP(0.8)
Pivot	Pivot - Most	0.74	0.79	0.64
	Pivot - Least	0.59	0.54	0.58
	Most - Least	0.66	0.58	0.63
	Max - Pivot	0.89	0.90	0.88
	Min - Pivot	0.87	0.93	0.84
Most	Pivot - Most	0.51	0.80	0.57
	Pivot - Least	0.41	0.63	0.42
	Most - Least	0.36	0.57	0.33
	Max - Most	0.82	0.90	0.84
	Min - Most	0.64	0.88	0.62
Least	Pivot - Most	0.67	0.88	0.66
	Pivot - Least	0.44	0.72	0.42
	Most - Least	0.60	0.77	0.53
	Max - Least	0.87	0.87	0.85
	Min - Least	0.90	0.86	0.90

to inform the understandability-biased evaluation and compute uRBP and its graded version, uRBPgr [13]. The analysis proceeds on a similar path to that in the previous section about topical relevance assessments.

5.1 Inter-Assessor Agreement among Paid Assessors

We analyse the understandability assessments for the queries for which assessments were collected from two paid assessors; we further use the Max_Label and Min_Label from Section 4.1 to combine labels. To compute understandability-biased measures, we use the topical relevance assessments from the Default set (the original CLEF 2015 labels). Statistics about the amount of disagreement between paid assessors in terms of assessments of understandability are reported

Table 5: Kendall- τ rank correlation for comparison of system ranking for understandability measures.

Section	Topical Set	Understandability Set	uRBP(0.8)	uRBPgr(0.8)
Section 5.1	Default	Default - Max_Scores	0.91	0.96
	Default	Default - Min_Scores	0.97	0.98
Section 5.2	Default44	Default44 - ICS	0.82	0.85
	ICS	ICS - Default44	0.83	0.86
	Default44	Layperson - Default44	0.82	0.87
	Default44	Layperson - ICS	0.85	0.90
Section 5.3	Default	Default - Pivot	0.77	0.75
	Default	Default - Most	0.74	0.71
	Default	Default - Least	0.72	0.71

in Table 2 and, overall, demonstrate similar levels of disagreement between assessors as for the topical relevance labels. Correlations between system rankings are reported in Table 5. Regardless of the specific label aggregation method and understandability measure, there is high correlation between system rankings produced with differing understandability assessments, suggesting system rankings are stable despite assessor disagreements. This is in line with the findings reported in Section 4.1 for topical relevance assessments.

5.2 Influence of Assessor Type and Payment Level

Next, we study differences due to assessor type (medical student vs. layperson) and payment level (paid vs. unpaid medical students). Table 2 reports the disagreements between the group of unpaid medical students and the 5 paid students, when assessing understandability. Overall they demonstrate close mean assessments, with the largest differences occurring due to Assessor 2, who tended to assess documents with a stricter view about understandability. The smallest differences are instead observed due to Assessor 4; however this did not show statistically significant differences ($p > 0.05$). Disagreement statistics among laypeople are available as an online appendix; they show a similar trend to those for unpaid students.

Table 5 reports the correlations between system rankings obtained when using the Default assessments and those with the ICS group and the laypeople group. These results demonstrate that, regardless of who performs the understandability assessments, high correlations across values of understandability-biased measures are obtained. This suggests that the use of either unpaid medical students or unpaid laypeople to assess understandability in place of paid medical students does not negatively influence the reliability of system rankings and evaluation. Thus, neither payment levels nor expertise influence the abilities of assessors to judge understandability: while there are assessment disagreements, these have limited impact on evaluation. This is unlike the results obtained in Section 4.2 when examining topical relevance.

5.3 Assessors Agreement across Query Variations

Here we study how understandability assessments vary across query variations for the same information need and what is the impact of potential disagreements

on the evaluation based on understandability-biased measures. Figure 2 reports the uRBP values (with $\rho = 0.8$) for each system across the three types of query variations (Pivot, Most and Least); Table 5 lists the correlations between each type of query variation and the default system ranking. Results here are similar to those obtained when investigating topical relevance (see Section 4.3) and support the importance of query variations for system evaluation.

6 Conclusions

In this paper we have examined assessment agreement between annotators across a number of different facets, including domain expertise, payment level, query variations, and assessment type (i.e., topical relevance and understandability).

Our analysis shows that there are often assessment disagreements both among assessors of the same type (e.g., among paid medical students) and among assessors of different types (e.g., among paid and unpaid medical students). Neither payment level, nor domain expertise and assessment type had significant influence in reducing the amount of disagreement across assessors.

We show that while assessor disagreement within the same type of assessor does not influence system rankings and evaluation, assessor disagreement with respect to topical relevance across types of assessors lead to lower correlations between system rankings. This results in unreliable system comparisons and thus evaluation if unpaid assessors or assessors with lower expertise are used in place of gold (paid, expert) assessors. This finding confirms results of previous research [2,4]. However, we also show that this is not the case when assessments of understandability, rather than of topical relevance, are sought. Our results in fact demonstrate that correlations between system rankings obtained with understandability-biased measures are high, regardless of payment levels and expertise. This is a novel finding and suggests that (1) Laypeople understandability assessments of health information on the web can be used in place of those of experts; and (2) The adoption of a two-stage approach to gather multi-dimensional relevance assessments where assessments are gathered from different types of assessors (both due to payment and expertise) may be viable, in particular if the assessment of dimensions beyond topicality requires additional time. In the first stage of such a method, assessor time from highly-paid, expert assessors is focused on assessing topical relevance. Labels produced by these assessments are to be used as a basis for both topical relevance measures (P@10, MAP, RBP, etc.) and understandability-biased measures. In the second stage, understandability assessments are acquired employing less expert or less expensive assessors, e.g., laypeople or through inexpensive graduate in-class activities. The use of such a two-stage approach for collecting assessments has the potential of reducing the overall cost of evaluation, or, with a fixed certain assessment-budget, of allowing to assess more documents. In addition, this approach may reduce the implicit dependencies assessors have between judging the different dimensions of relevance.

Finally, our results add to the recent body of work showing the importance of query variations for increasing the reliability and veracity of Informa-

tion Retrieval evaluation [1,2]. We show, in fact, that the availability of query variations for an information need contribute great diversity to the pool and that system rankings obtained with only one of the three types of variation considered here are unstable when compared with the rankings obtained with all variations (both for topical relevance and understandability). The data and code used in this research is available online at <https://github.com/ielab/clef2016-AssessorAgreement>.

Acknowledgements This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect), and from the Austrian Science Fund (FWF) projects P25905-N23 (ADmIRE) and I1094-N23 (MUCKE).

References

1. L. Azzopardi. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proc. of SIGIR*, pages 556–563, 2009.
2. P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does It Matter. In *Proc. of SIGIR*, pages 667–674, 2008.
3. P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User Variability and IR System Evaluation. In *Proc. of SIGIR*, pages 625–634, 2015.
4. B. Carterette and I. Soboroff. The Effect of Assessor Error on IR System Evaluation. In *Proc. of SIGIR*, pages 539–546, 2010.
5. B. Koopman and G. Zuccon. Relevation!: an open source system for information retrieval relevance assessment. In *Proc. of SIGIR*, pages 1243–1244. ACM, 2014.
6. B. Koopman and G. Zuccon. Why assessing relevance in medical IR is demanding. In *Medical Information Retrieval Workshop at SIGIR 2014*, 2014.
7. M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.
8. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2, 2008.
9. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information About Medical Symptoms. In *CLEF*, 2015.
10. I. Stanton, S. Jeong, and N. Mishra. Circumlocution in diagnostic medical queries. In *Proc. of SIGIR*, pages 133–142. ACM, 2014.
11. E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.
12. E. M. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval*, volume 1. MIT press Cambridge, 2005.
13. G. Zuccon. Understandability biased evaluation for information retrieval. In *Advances in Information Retrieval*, 2016.
14. G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Medical Information Retrieval Workshop at SIGIR 2014*, page 32, 2014.
15. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in Information Retrieval*, pages 562–567. Springer, 2015.