The University Of Queensland
AUSTRALIA
CREATE CHANGE

Better Search, Better Health?

Search Engines, their Evaluation and the Impact on Health Decisions

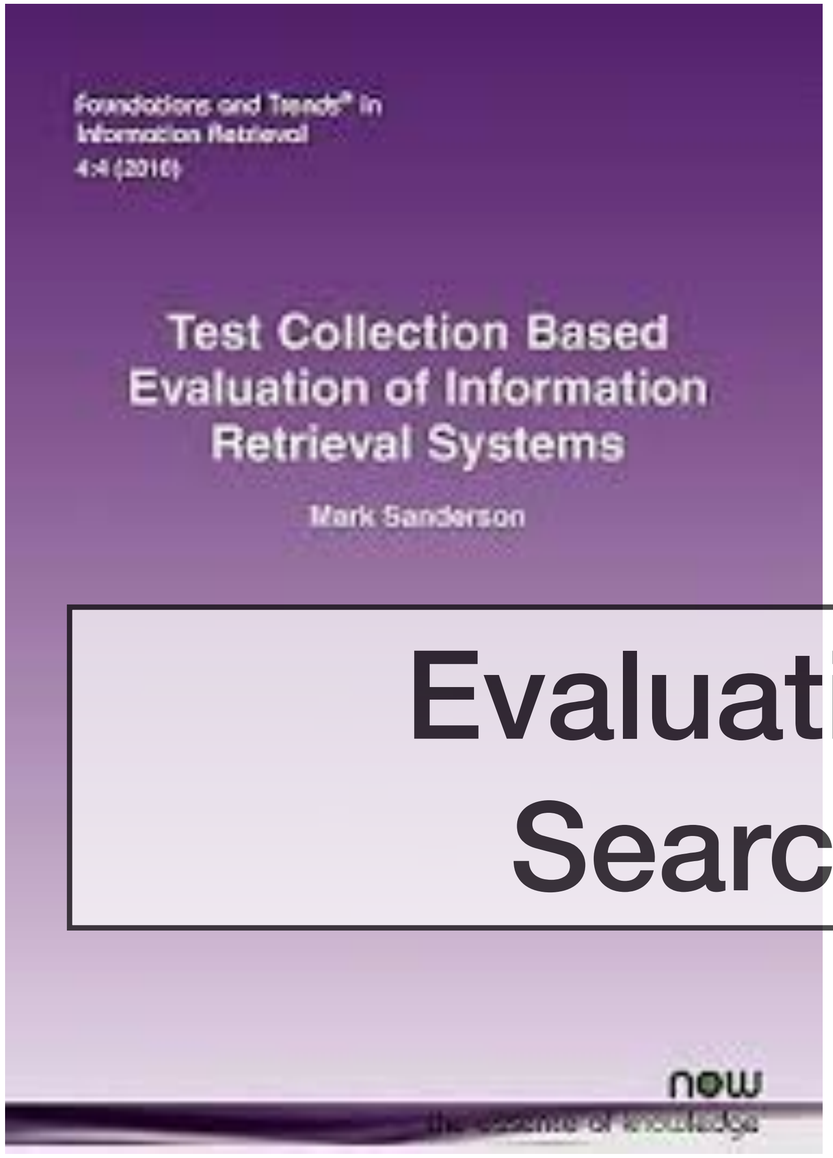
Guido Zuccon

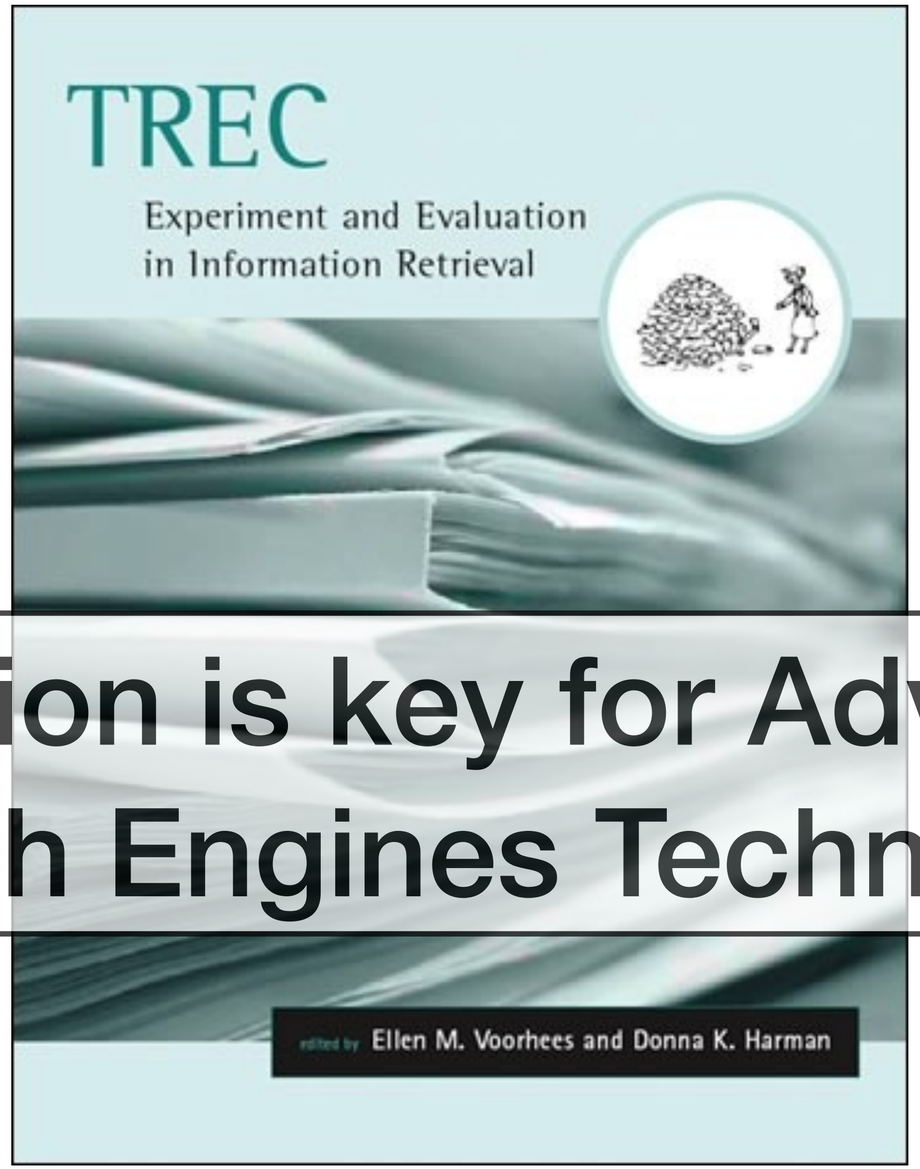
g.zuccon@uq.edu.au — ielab.io/guido

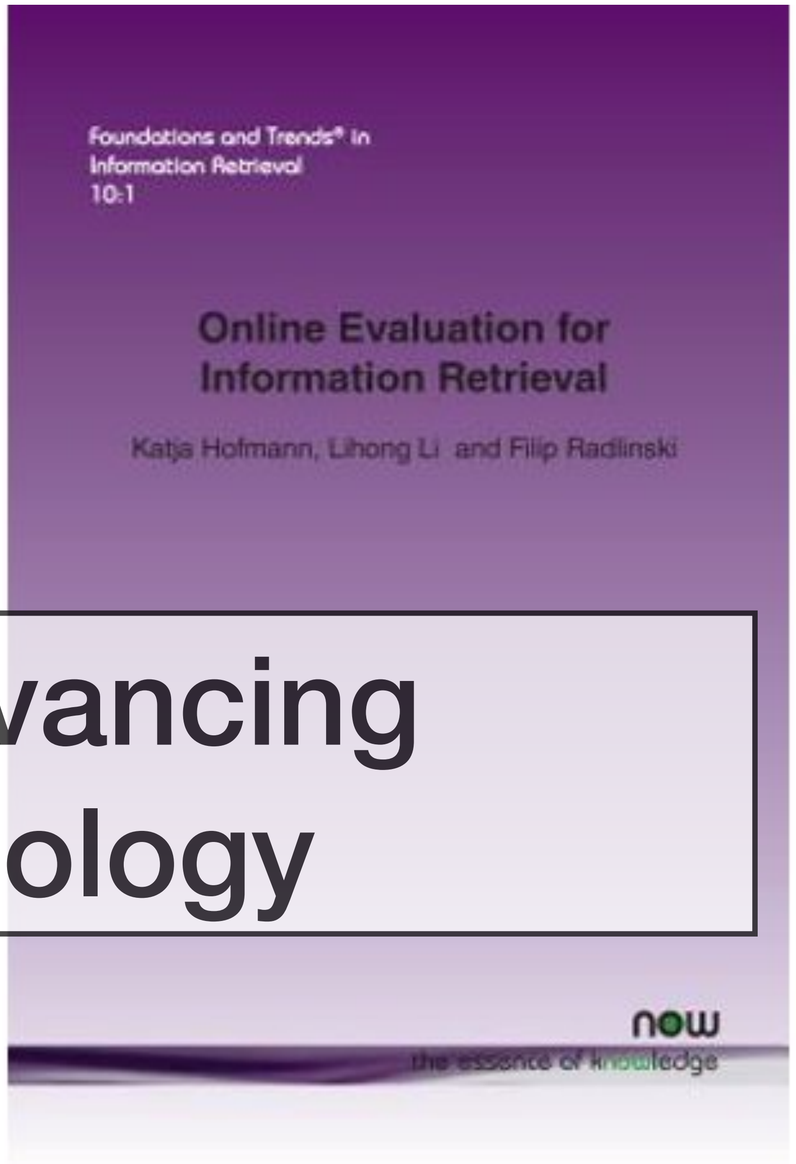
The University of Queensland

ie lab

Work presented here is in collaboration with: Anton van der Vegt, Jimmy, Bevan Koopman (AEHRC/CSIRO), Gianluca Demartini, Leif Azzopardi (Strathclyde)







Evaluation is key for Advancing Search Engines Technology

IR evaluation focuses on how much relevant material the system retrieves*,**

Is this all we care about?

*** In some user studies, user satisfaction and perceived task completion are also probed**
**** In online evaluation, clicks are used as an indicator of relevance/engagement**

Why do we search?

- Broder 2002's taxonomy of search intents: **navigational** (20% - reach a known site), **informational** (~50% - acquire information), **transactional** (30% - perform activity)

- We often do this for **completing a task**

- Emergent research on supporting users in **completing tasks**

- **Task completion support technology**, e.g.
  - Mehrotra & Yilmaz, "Task embeddings: Learning query embeddings using task context" CIKM 2017.
  - Koopman, Zuccon, Russell. "A Task-oriented Search Engine for Evidence-based Medicine" SIGIR 2017.

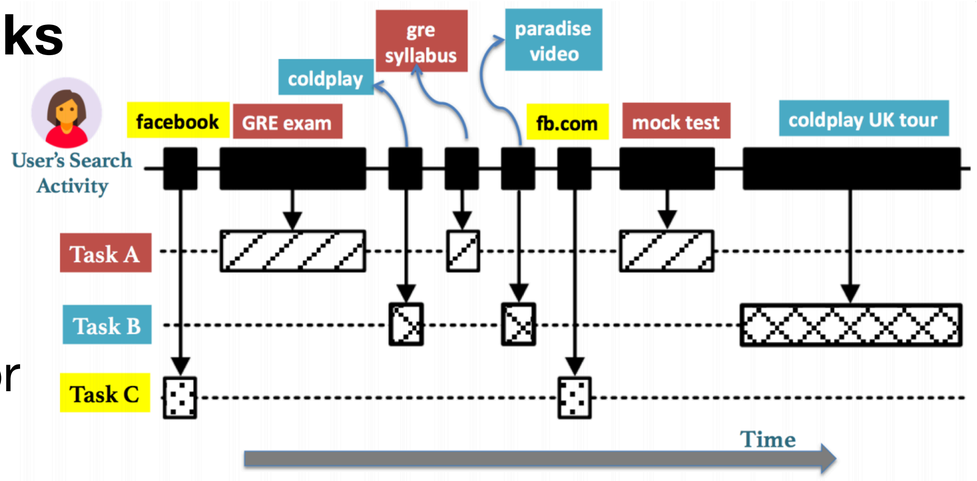


Image from Mehrotra & Yilmaz, CIKM2017

- **Task-based evaluation**, e.g.
  - Verma et al. "Overview of the TREC Tasks Track 2016" TREC 2016.
  - Lipani, Carterette, Yilmaz. "From a User Model for Query Sessions to Session Rank Biased Precision (sRBP)" SIGIR 2019.

Searching to make decisions

We go one step forward: investigate situations where we **search to make a** (critical) **decision**

In this talk:

Do search engines allow us to make better decisions?

- I will reflect on the **implications** of our empirical results and observations with respect to:
  - How we **evaluate** search systems in Information Retrieval
  - What **future research** may need to address

1st Study: Clinical Decision Support

Clinical Questions: An information need that arises as a result of a patient encounter directly pertaining to the diagnosis and/or treatments for a patient condition

- What is the **diagnosis**?
- What **medical tests** to order?
- What **treatments** to suggest?

- 1 question every 2 patient visits
- More than half (53%) of clinical questions don't get answered

[Del Fiol et al., 2014]

Does it matter if clinical questions are answered?

- **answering clinical questions** leads to significant changes to care

- relevant information provided to clinicians -> **changes to clinical decisions are observed** for:
    - Patient **advice** (72%)
    - **Treatment** decision (59%)
    - Choice of diagnostic **test** (45%)
    - Choice of drug (**treatment**) (45%)
    - **Diagnosis** decision (29%)

Marshall JG. The impact of the hospital library on clinical decision making: the Rochester study. Bull Med Libr Assoc. 1992;80(2):169.

Marshall JG, et al. The value of library and information services in patient care: results of a multisite study. J Med Libr Assoc JMLA. 2013;101(1):38.

- Thus **avoiding**: additional **tests** and **procedures** (49%), additional **surgery** (21%), patient **mortality** (19%)

- **Correctly answering** clinical questions **impacts** patient **outcomes** and the **efficiency** and **effectiveness** of the healthcare system

Does it matter if clinical questions are answered?

- **answering clinical questions** leads to significant changes to care

- relevant information provided to clinicians -> **changes to clinical decisions are observed** for:
  - Patient **advice** (72%)
  - **Treatment** decision (59%)
  - Choice of diagnostic **test** (45%)
  - Choice of drug (**treatment**) (45%)
  - **Diagnosis** decision (29%)

Marshall JG. The impact of the hospital library on clinical decision making: the Rochester study. Bull Med Libr Assoc. 1992;80(2):169.

Marshall JG, et al. The value of library and information services in patient care: results of a multisite study. J Med Libr Assoc JMLA. 2013;101(1):38.

Search Engines often used by clinicians for answering their clinical questions

- PubMed
- UpToDate

Does search engine lead to better clinical decisions?

- **User study** with 106 clinicians; 16 questions, 1,653 samples
- Clinicians asked to consider **real-life clinical scenarios** with patient complain description:
    - *A 48 year old man presents with severe right sided loin pain and is diagnosed with a 4mm distal ureteric calculus.*
- Clinician asked **focused question** on diagnosis/test/treatment
    - *Has Tamsulosin been shown to increase the chances of the calculus passing?*
- Answer = *Yes, No, Conflicting*
- Ask question **pre-search** — Ask again **post-search**

Based on: (1) van der Vegt A, Zuccon G, Koopman B, Deacon A. Impact of a Search Engine on Clinical Decisions Under Time and System Effectiveness Constraints: Research Protocol. JMIR Res Protoc 2019;8(5):e12803
(2) van der Vegt A, Zuccon G, Koopman B, Deacon A. The Impact of time pressure on clinical decision making. [under review]
(3) van der Vegt A, Zuccon G, Koopman B. Does better retrieval effectiveness equate to better clinical decisions? If not, why not? [under review]

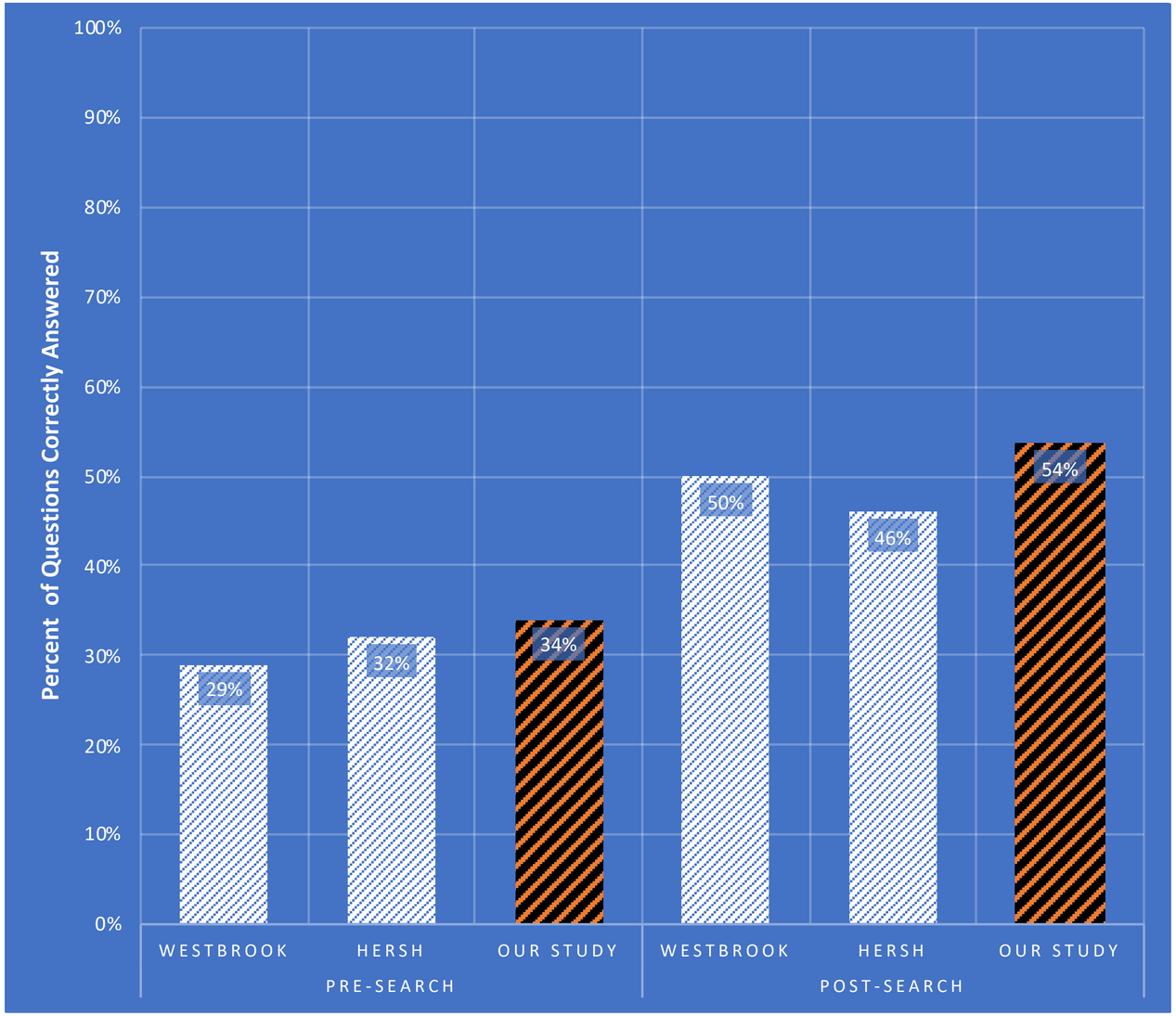
Does search engine lead to better clinical decisions?

Similar Prior Studies:

- [Westbrook et al, 2005]:
  75 clinicians, 8 clinical questions
- [Hersh et al, 2002]:
  45 medical students, 21 nursing students, total 324 questions

Our results:

- Slightly higher pre-search & post-search correct answer rate
- Similar **significant benefit** from using a medical **search** engine



Note: All differences between pre-search and post-search answers are significant (McNemar's Chi-squared Test for symmetry)

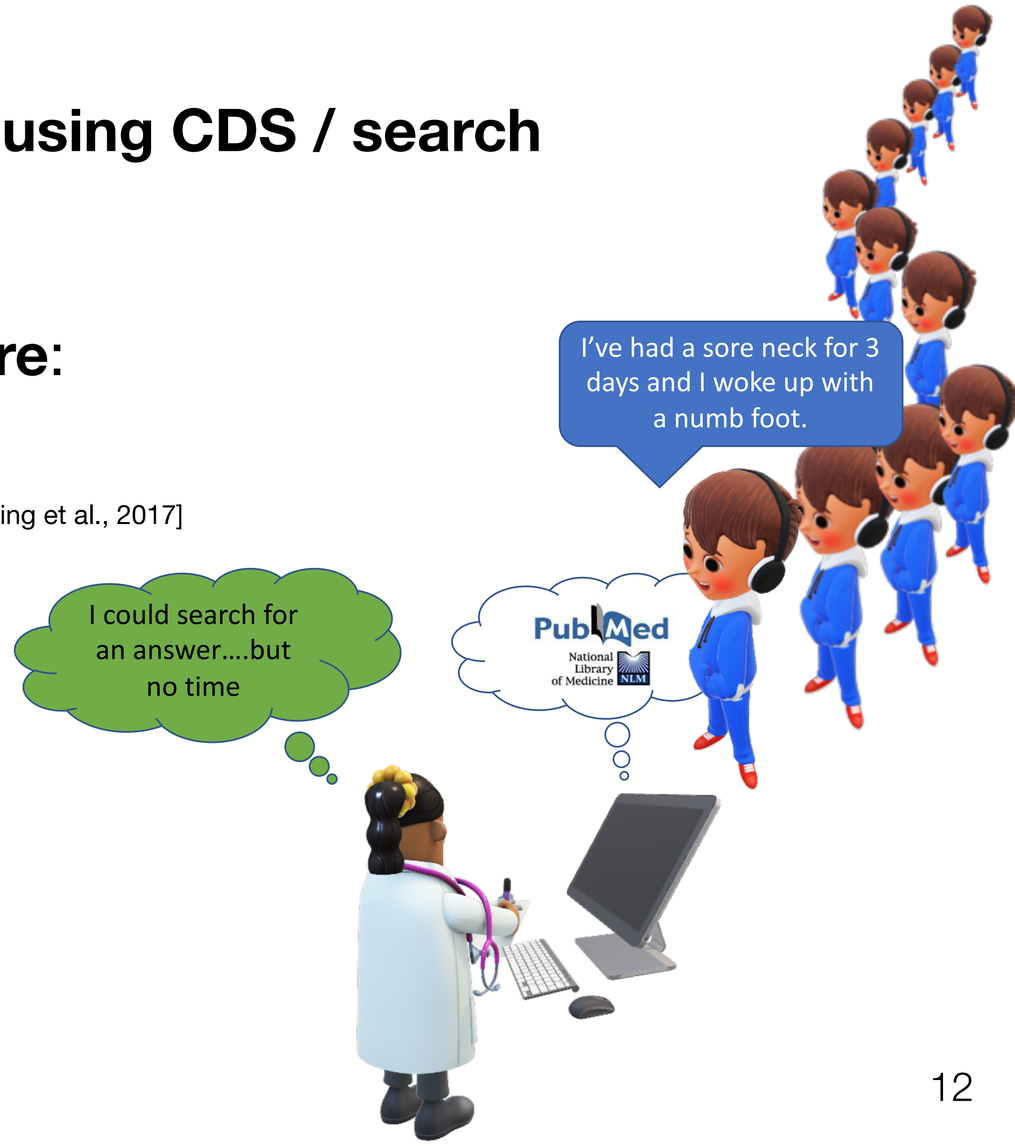
How does time pressure impact clinical decision making, when using a search engine?

- Clinicians report **lack of time** as the primary **barrier to using CDS / search systems** [Cook et al., 2013]

- In fact, clinicians frequently operate under **time pressure**:

  - Average consultation time is less than 5 minutes [Irving et al., 2017]

  - Over one-third of all primary care physicians are dissatisfied with the time available per patient [Osborn et al., 2015]

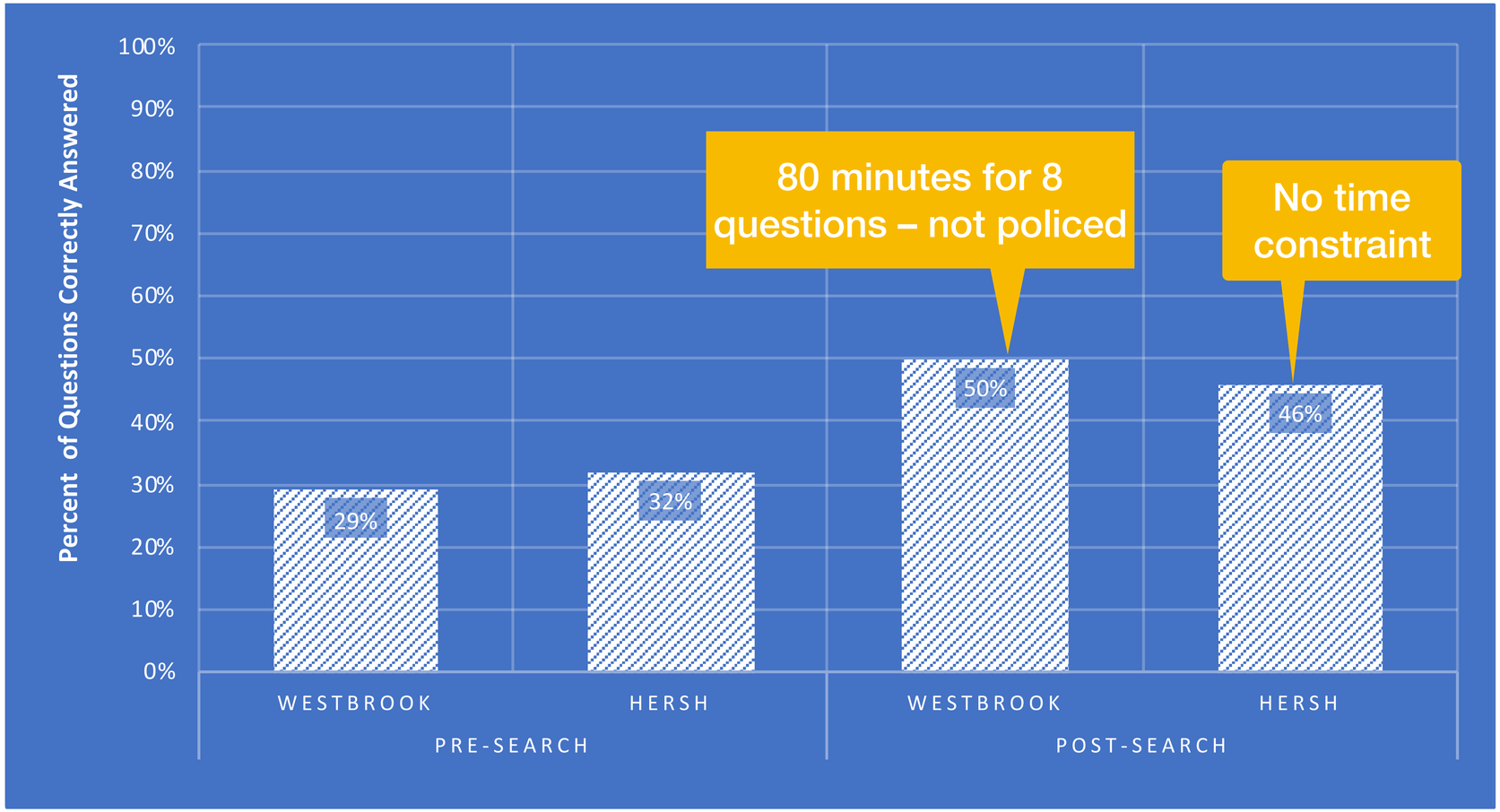


How does time pressure impact clinical decision making, when using a search engine?

Similar Prior Studies: time pressure not controlled for, or measured

[Crescenzi et al, 2016]: Impact of time constraints on user during information search. Users report:

- Greater time pressure
- Higher perceived task difficulty
- Less satisfaction with their performance



Study Design: Imposing Time Pressure

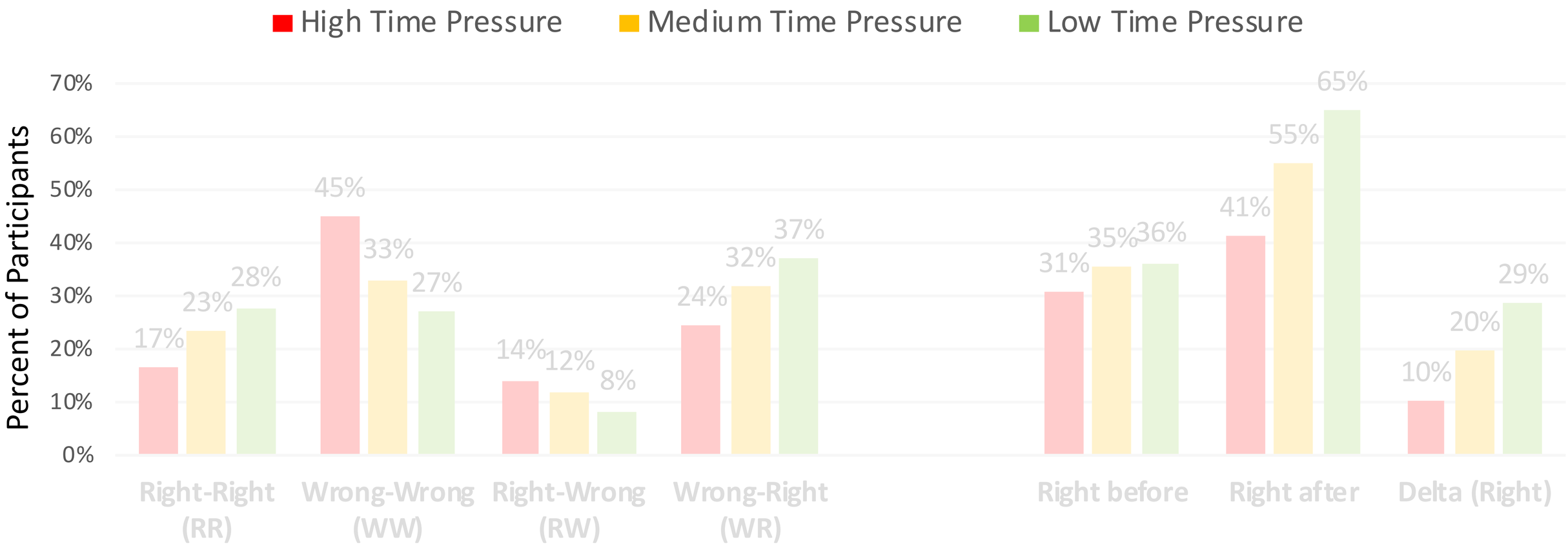
Search time constraints applied to induce time pressure

Within-subject, time constraint changes each question

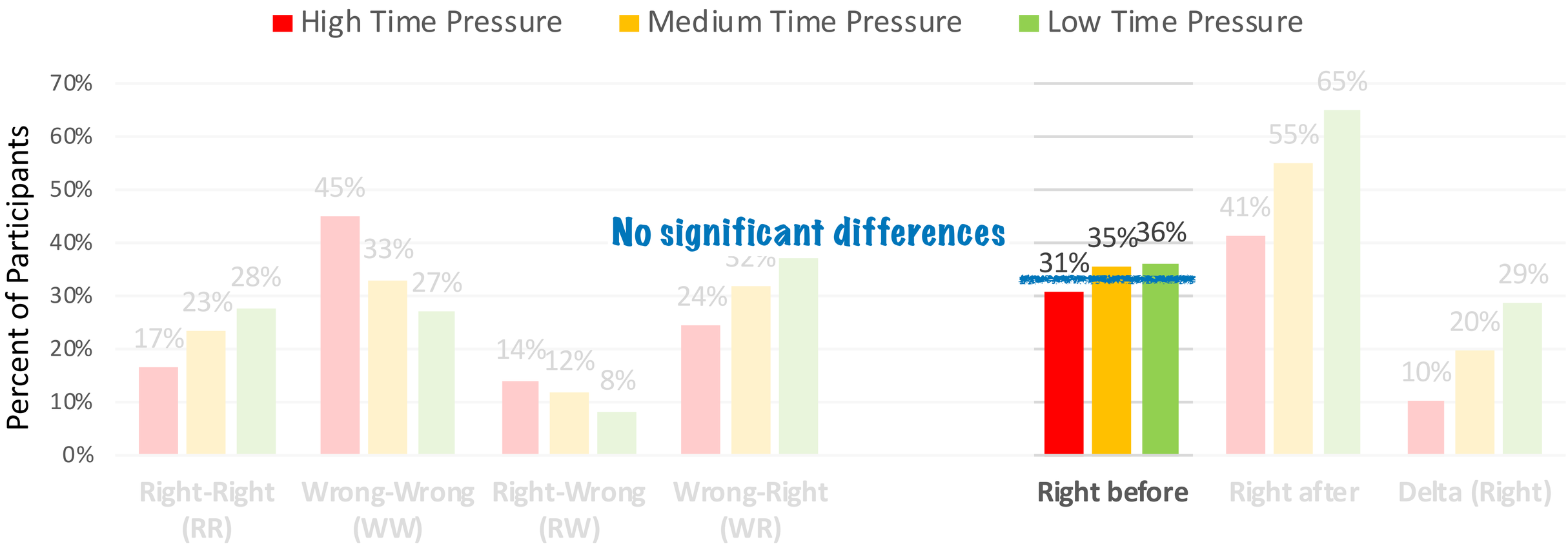
Based on [Westbrook et al, 2005]: mean search time = 6.1mins, std. dev = 3 mins. Therefore, 3 Time constraints applied to search:
3 minute, 6 minute and 9 minute

Participant notified of time at start of each question, then given a minute-by-minute countdown timer, then warning at 30 seconds to go, then forced to stop and answer question when time is up

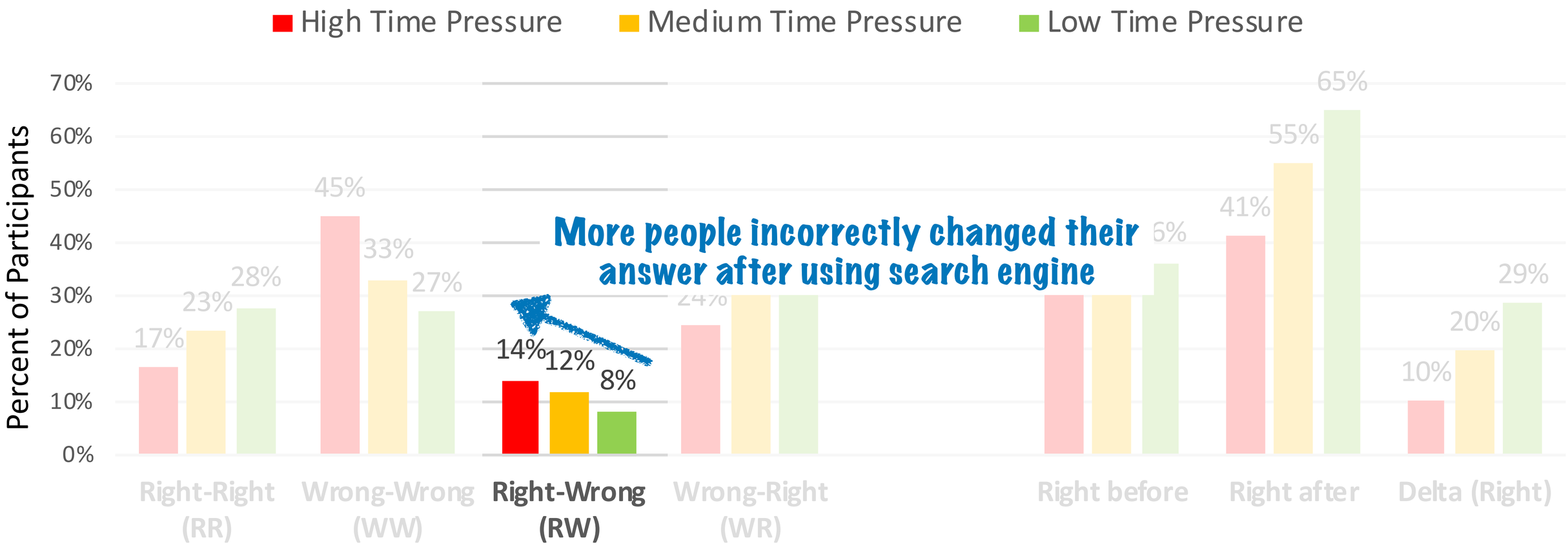
Impact of Time Pressure of Decisions



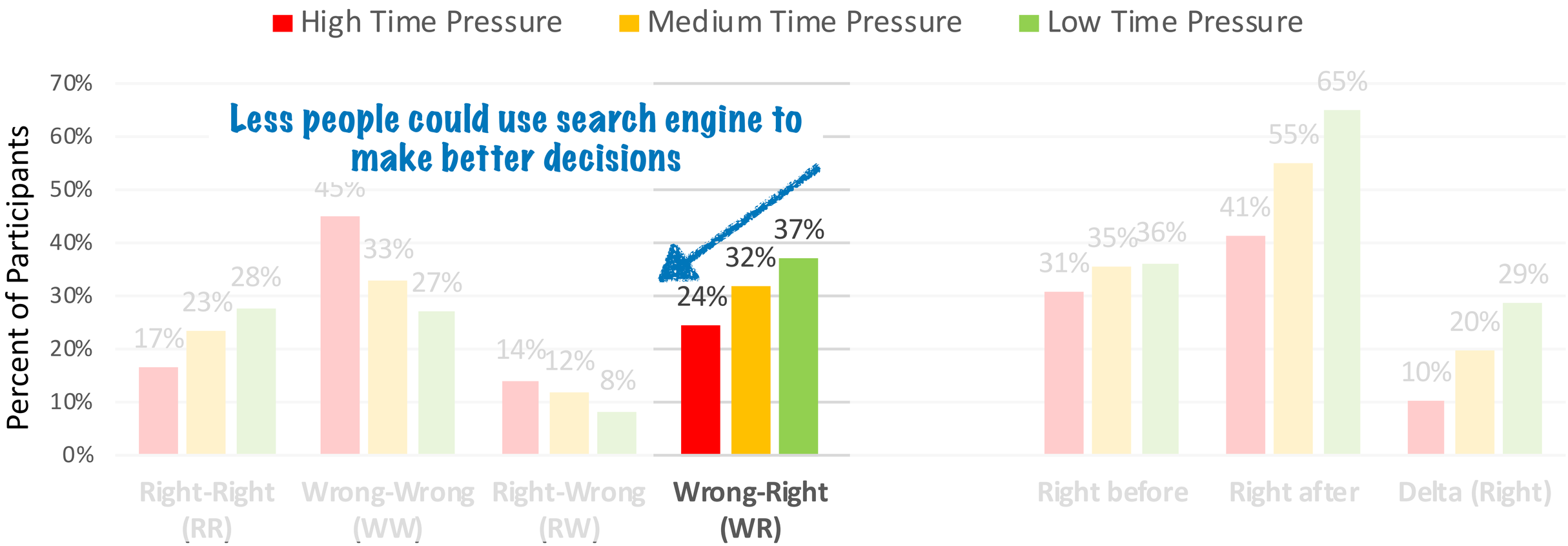
Correctness Before Search



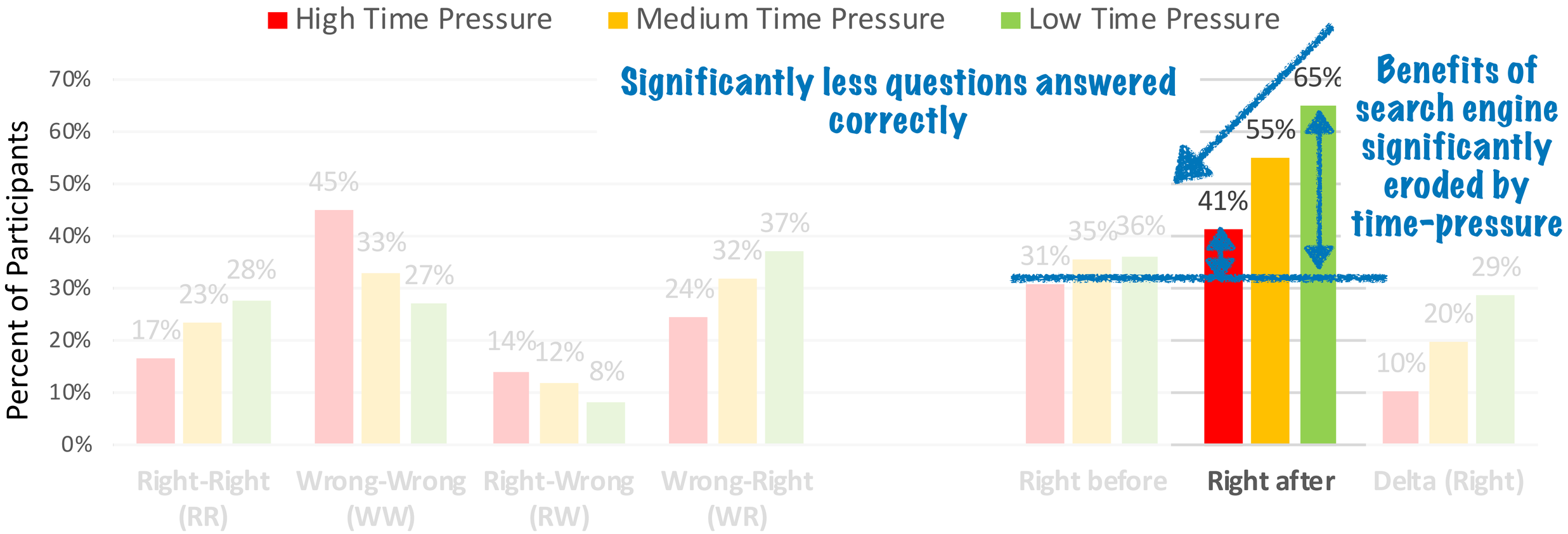
Right before search, then wrong after



Wrong before search then right



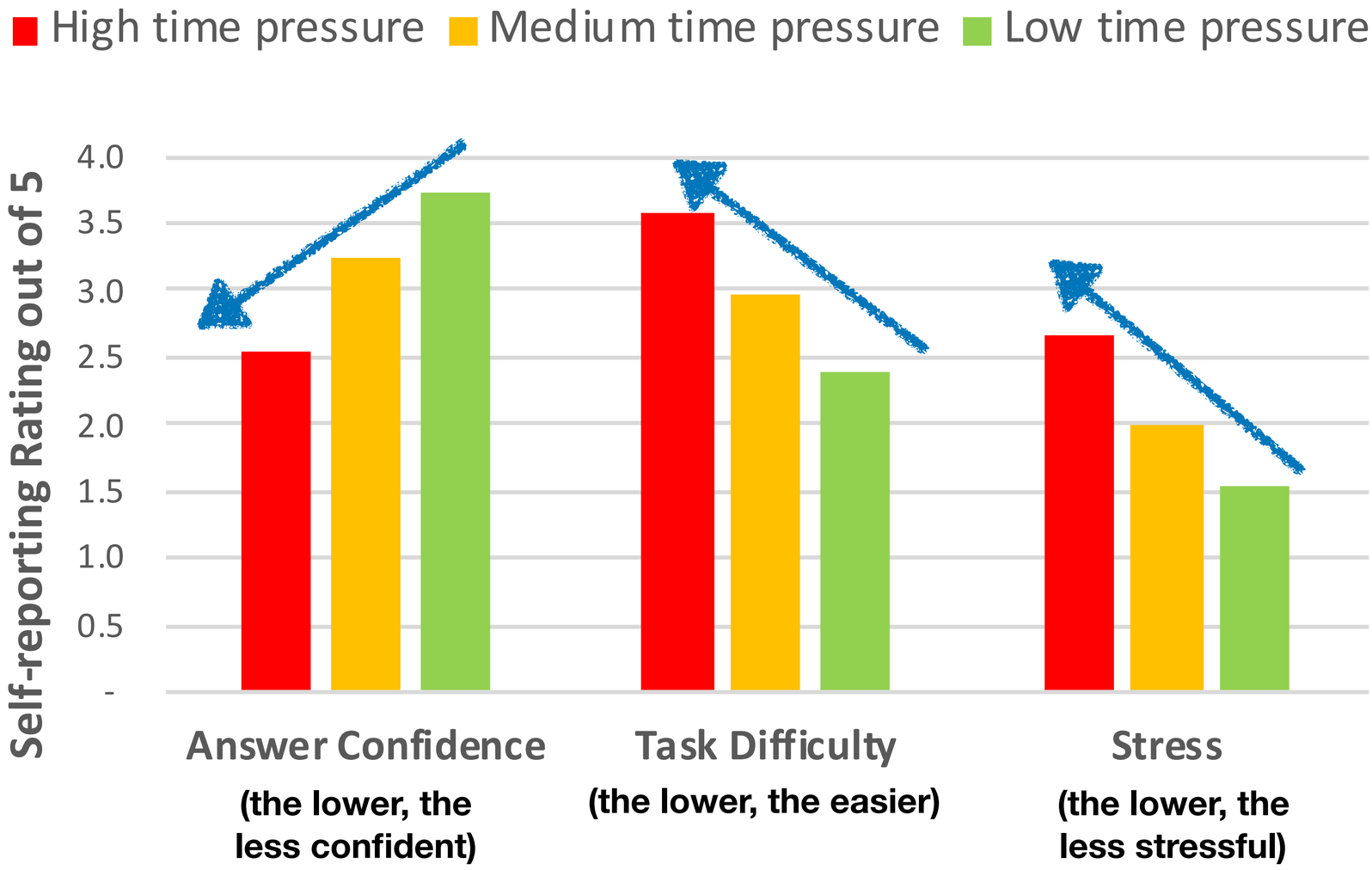
Right after using search system



Impact of time pressure on participants

- Decreased answer confidence
- Increased task difficulty
- Increased stress

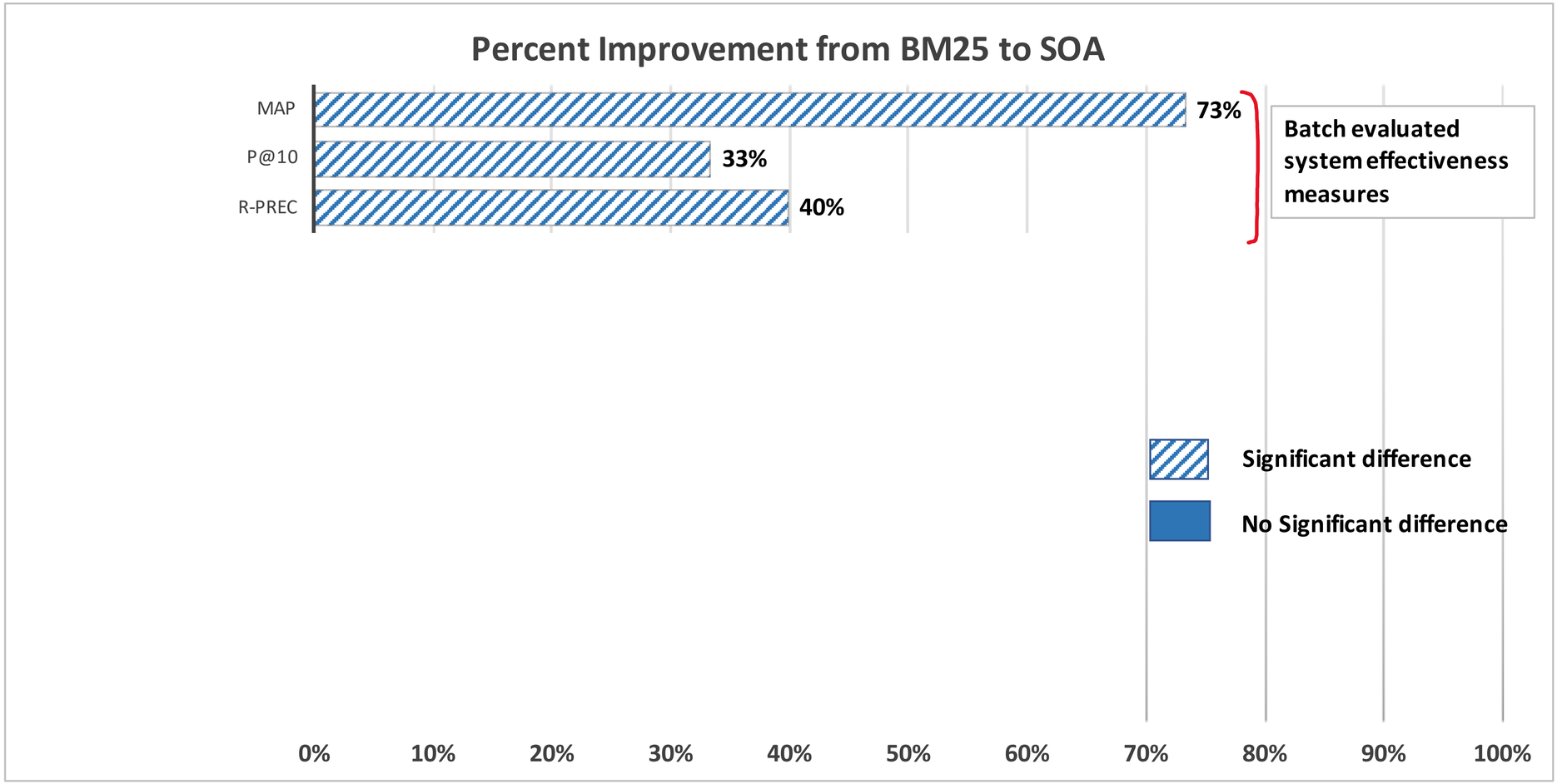
Note: All differences between time pressure levels are significant (TukeyHSD p<.05)



Does a significantly better search engine translate to better and faster clinical decisions?

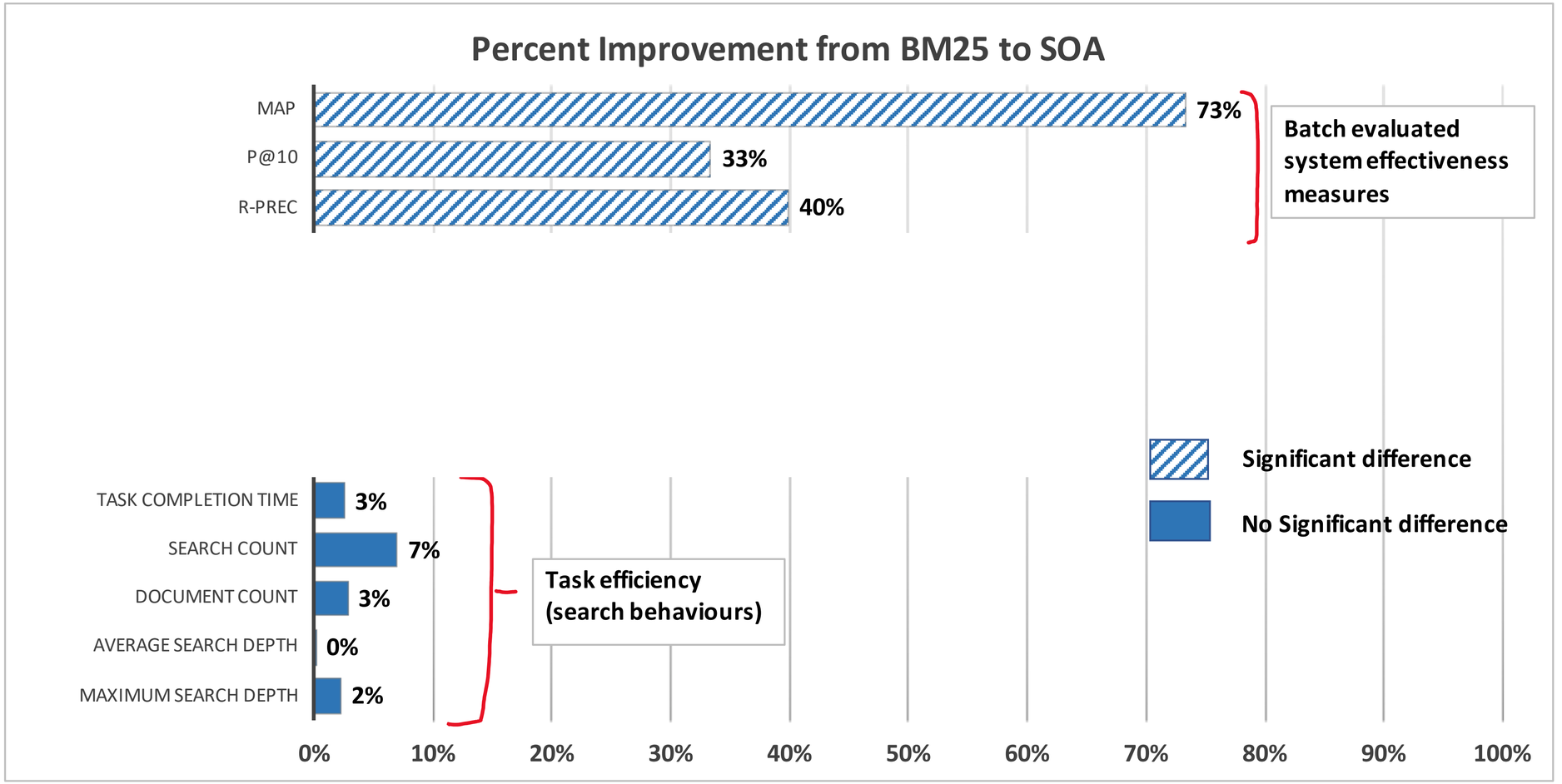
- After all, the IR community spends so much time and $$ **building better and better search engines** — in terms of good they are in **retrieving relevant** documents
    - e.g. for this task 7 TREC tracks (3x CDS, 4 x PM)
- In the user study we employed **two systems**
    1. Vanilla **BM25** (baseline)
    2. The state of the art (**SOA**) on the TREC CDS task: best system in 2015, further improvements in follow up SIGIR publications
        - Further tuned & improved RF pipeline, removed negated UMLS terms in query
- All systems tuned and validated on TREC CDS data

Does a significantly better search engine translate to better and faster clinical decisions?



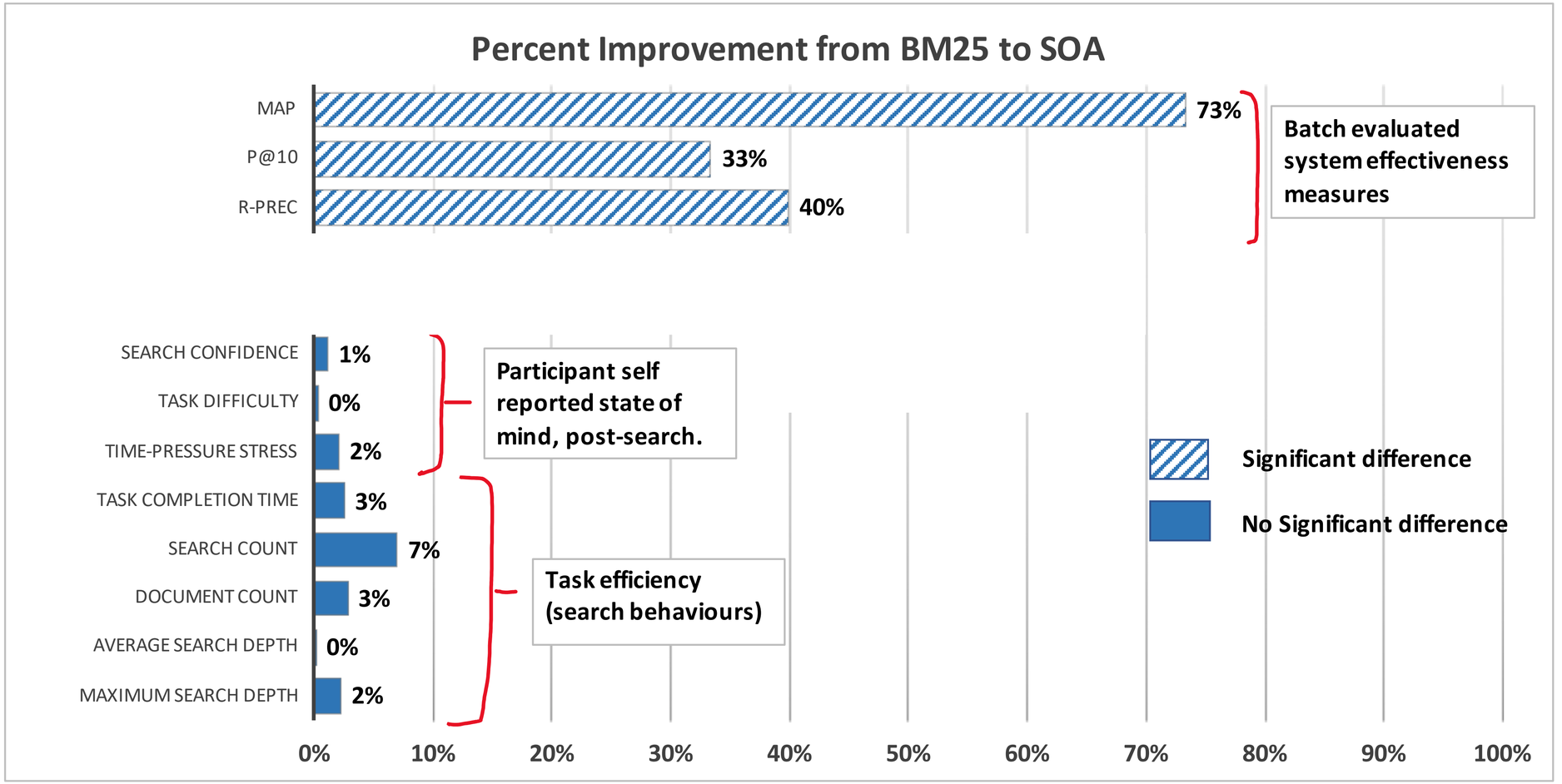
- Significant difference in off-line traditional IR performance
  - e.g., +73% MAP, +84% infAP, +46% infnDCG

Does a significantly better search engine translate to better and faster clinical decisions?



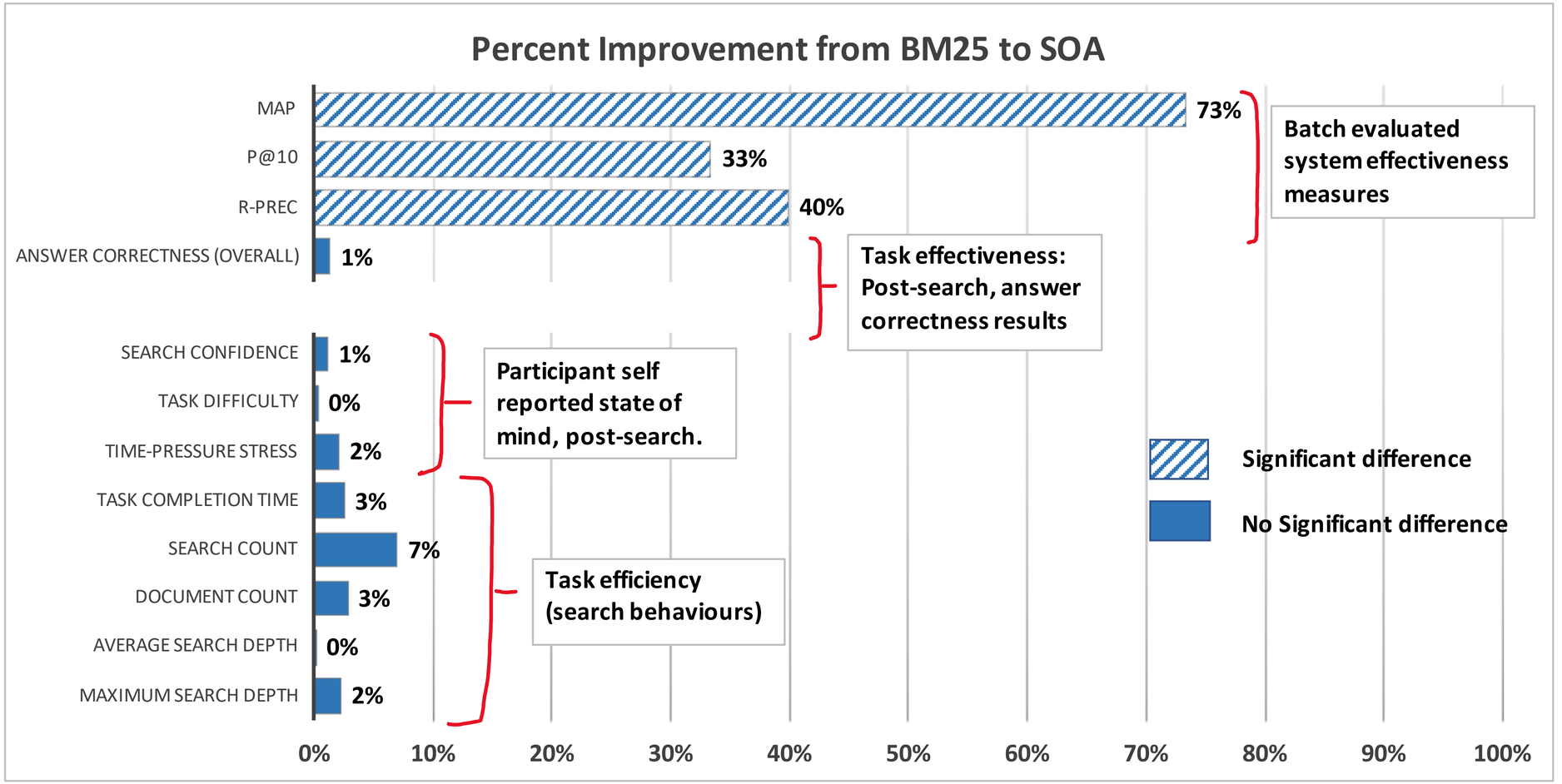
- Significant difference in off-line traditional IR performance
- **No/Minimal Impact on user search behaviour**

Does a significantly better search engine translate to better and faster clinical decisions?



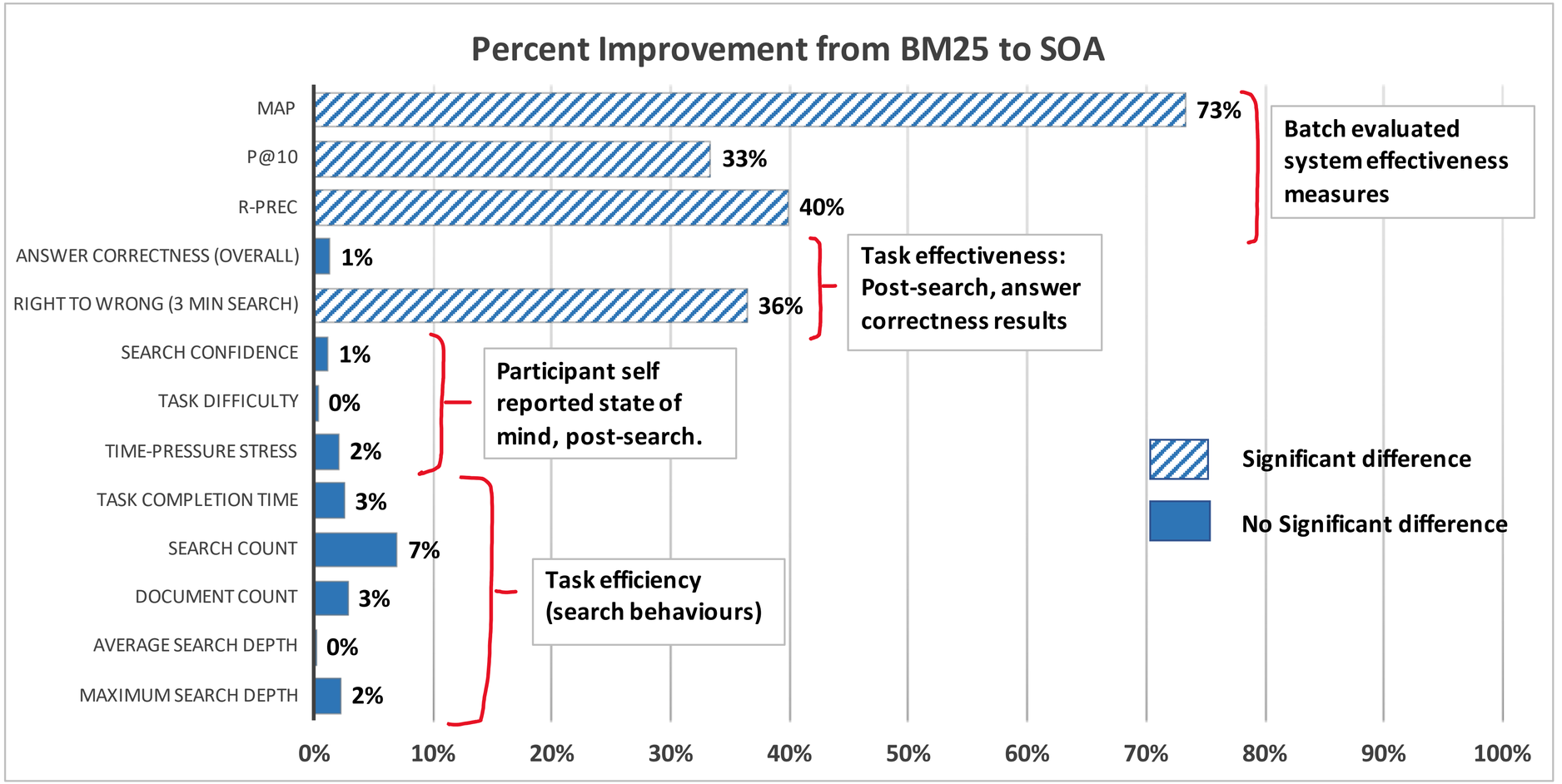
- Significant difference in off-line traditional IR performance
- No/Minimal Impact on user search behaviour
- **No/Minimal Impact on self-reported state of mind**

Does a significantly better search engine translate to better and faster clinical decisions?



- Significant difference in off-line traditional IR performance
- No/Minimal Impact on user search behaviour!
- No/Minimal Impact on self-reported state of mind
- **No/Minimal Impact on decisions**

Does a significantly better search engine translate to better and faster clinical decisions?

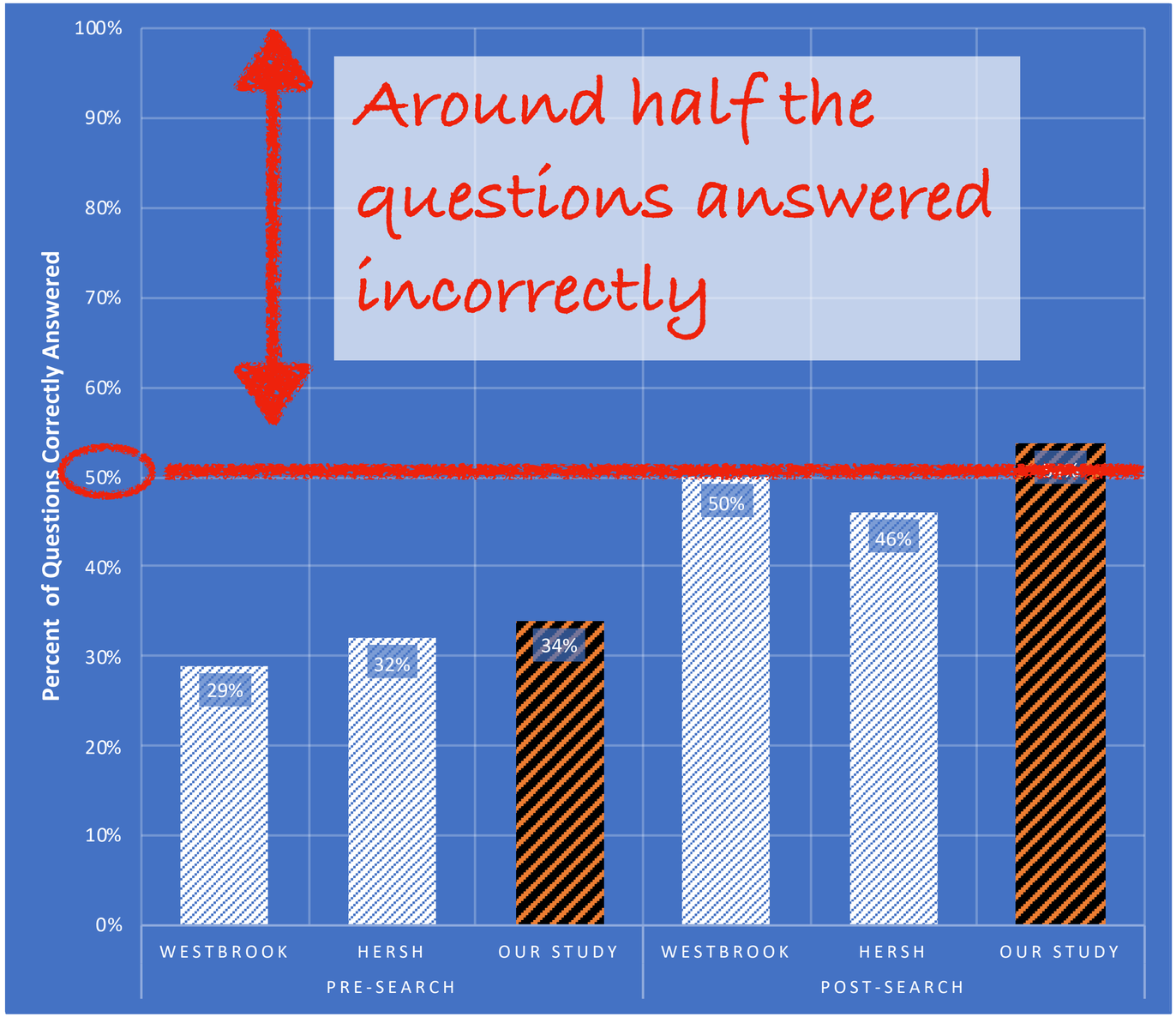


- Significant difference in off-line traditional IR performance
- No/Minimal Impact on user search behaviour!
- No/Minimal Impact on self-reported state of mind
- **No/Minimal Impact on decisions!**
  - Under the tightest, time-constrained search conditions, a better performing search engine **prevents** more **clinicians from incorrectly changing** their clinical **decisions**

Broader Implications for IR Evaluation

- Traditionally, **search engine evaluation** has measured a **function** of the presence of **relevant documents** in the ranking, e.g. precision, recall, nDCG, ~~MAP~~*, RBP*, INST*
- BUT — users do not search to find relevant documents
- They **search to fulfil tasks**, and often to **make decisions**
- Our results show that producing **good rankings** in terms of relevance does **NOT mean** helping people making **correct decisions**
- In cases such the one seen here, **evaluation should be more holistic**: it should consider the root cause for using the search engine, i.e. making decisions
- (see next for what may actually impact decisions)

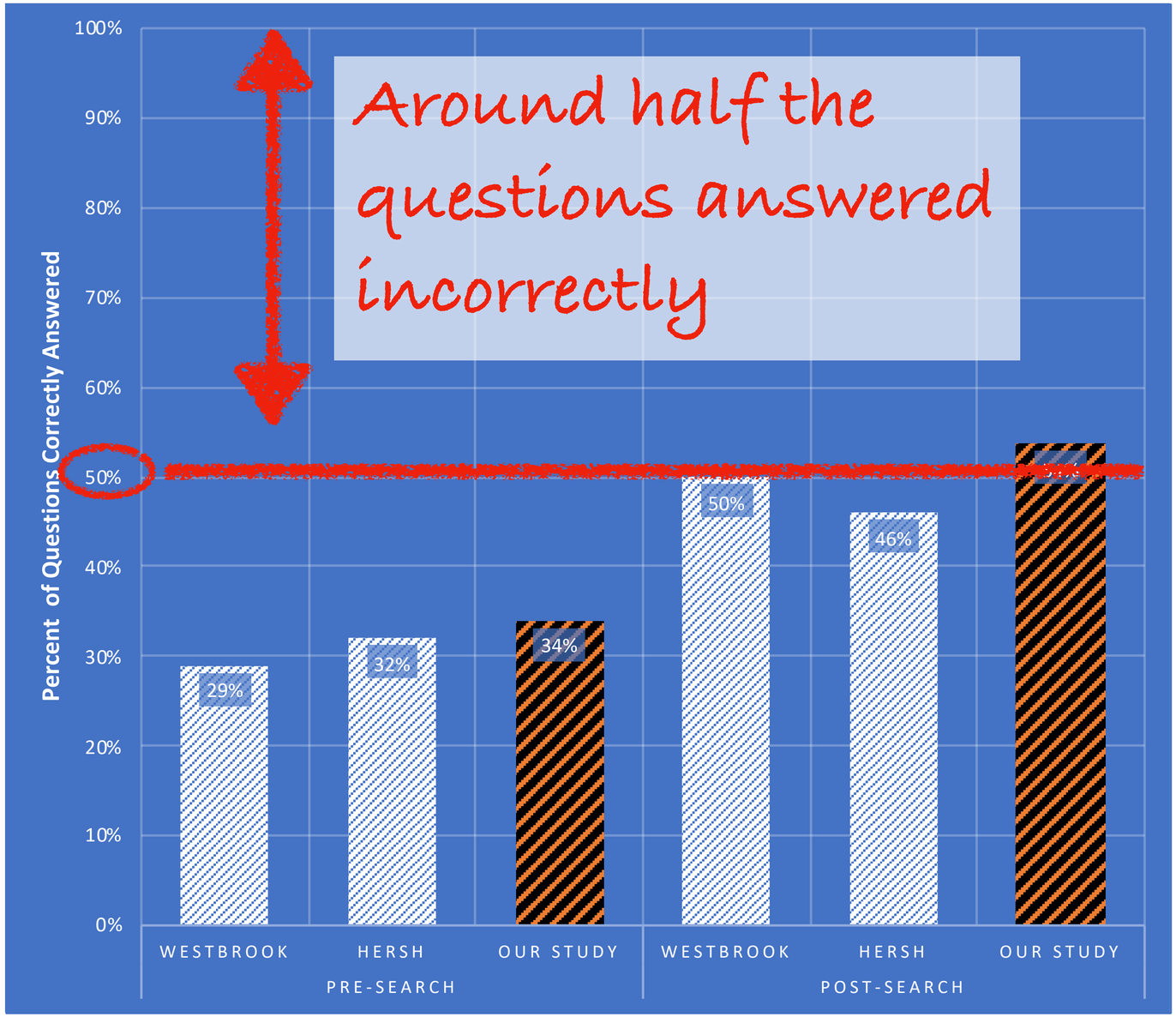
* Inserted just for Alistair to check if he is paying attention


Around half the questions answered incorrectly
Percent of Questions Correctly Answered
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
29%
32%
34%
50%
46%
WESTBROOK
HERSH
OUR STUDY
WESTBROOK
HERSH
OUR STUDY
PRE-SEARCH
POST-SEARCH

So, why do clinicians fail to correctly answer around half the questions?

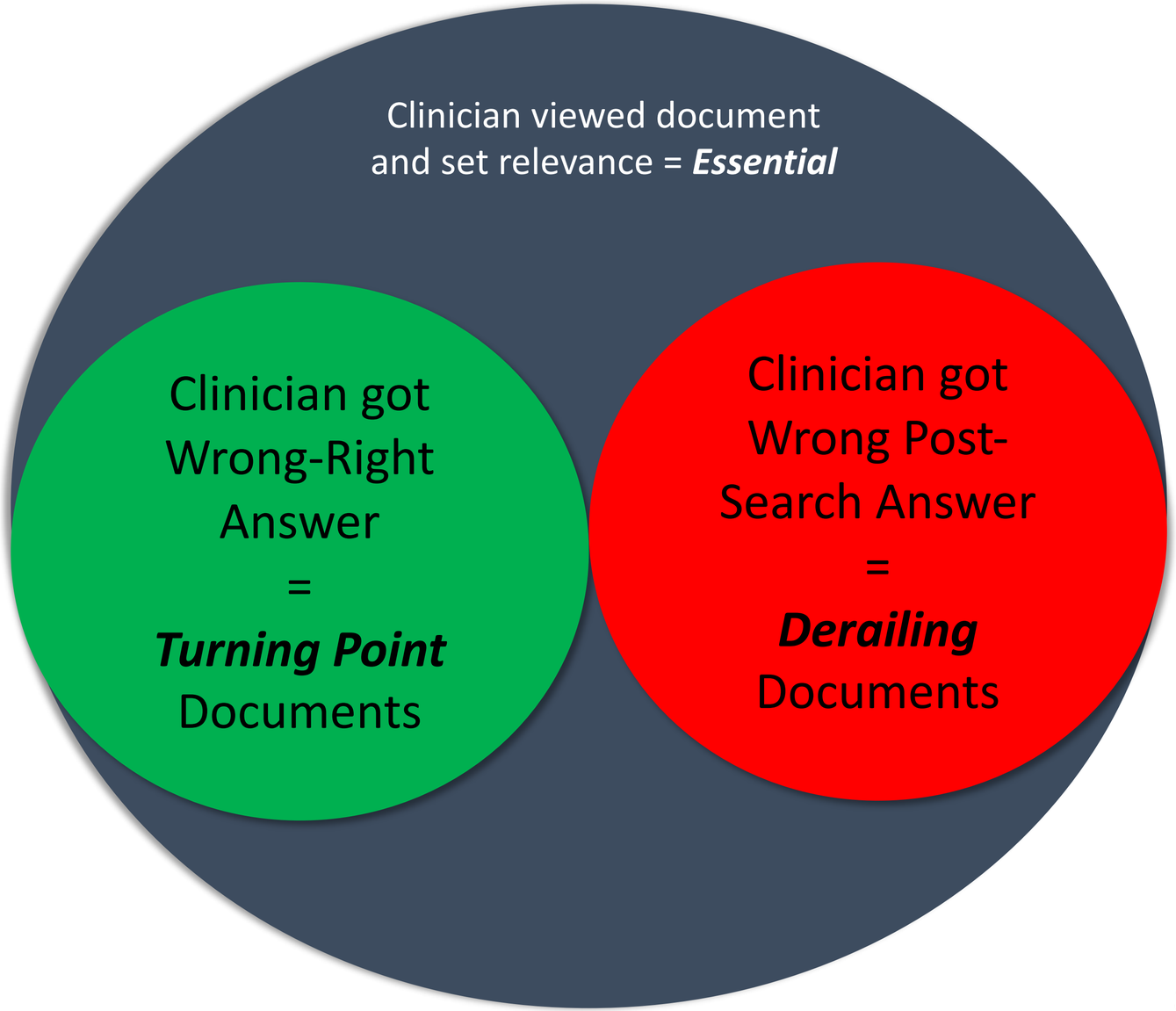
Is it because of:

- the corpus?
- the search engine?
- the participant?

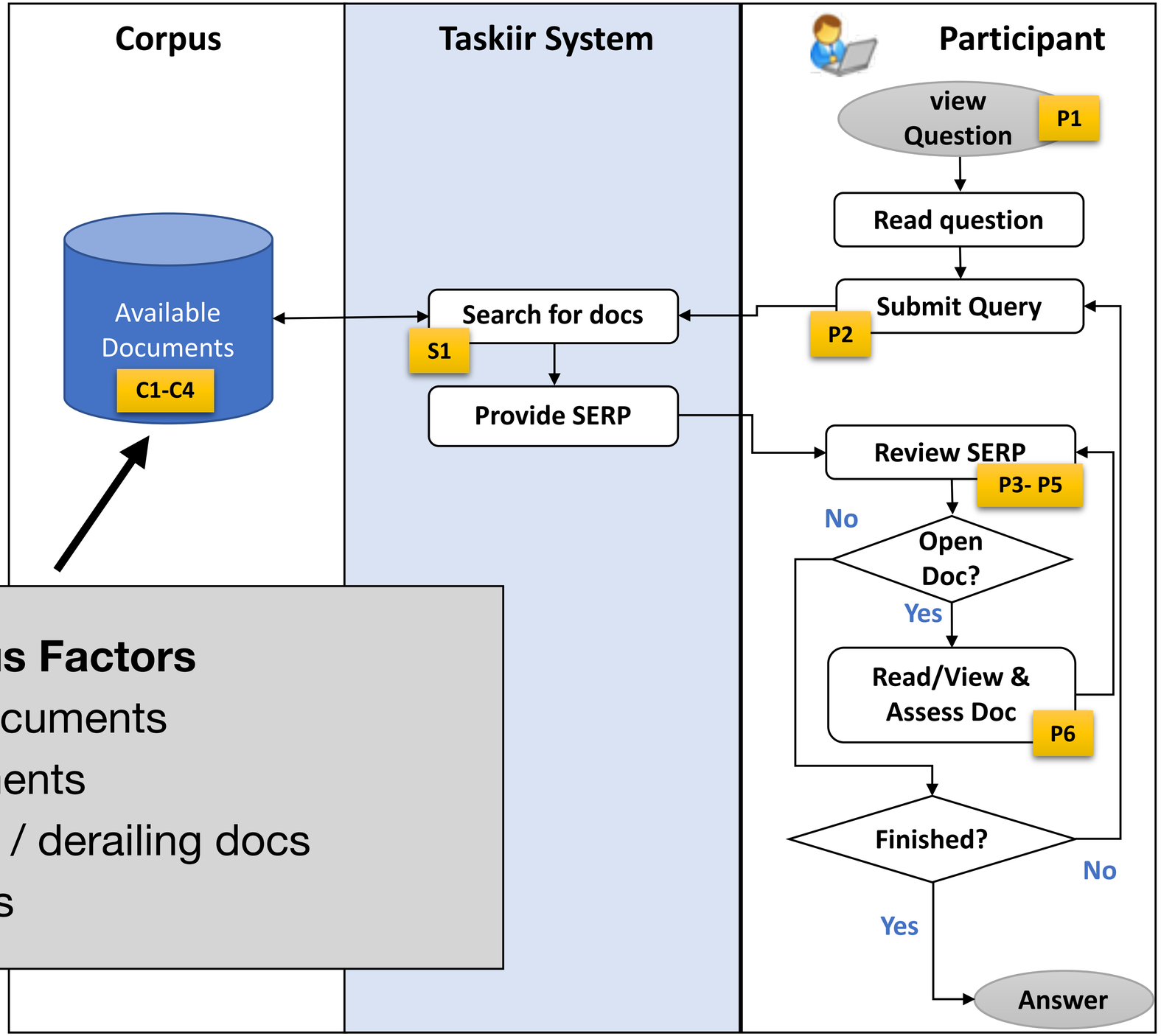


Analysis: Success Factor Model

- Three types of documents:
  - **Turning Point**: those identified essential and that lead to change answer from wrong to right
  - **Derailing**: those identified essential but that lead to wrong answer
  - **Ambiguous**: those identified as both turning point and derailing



The Success Factors



Corpus Factors

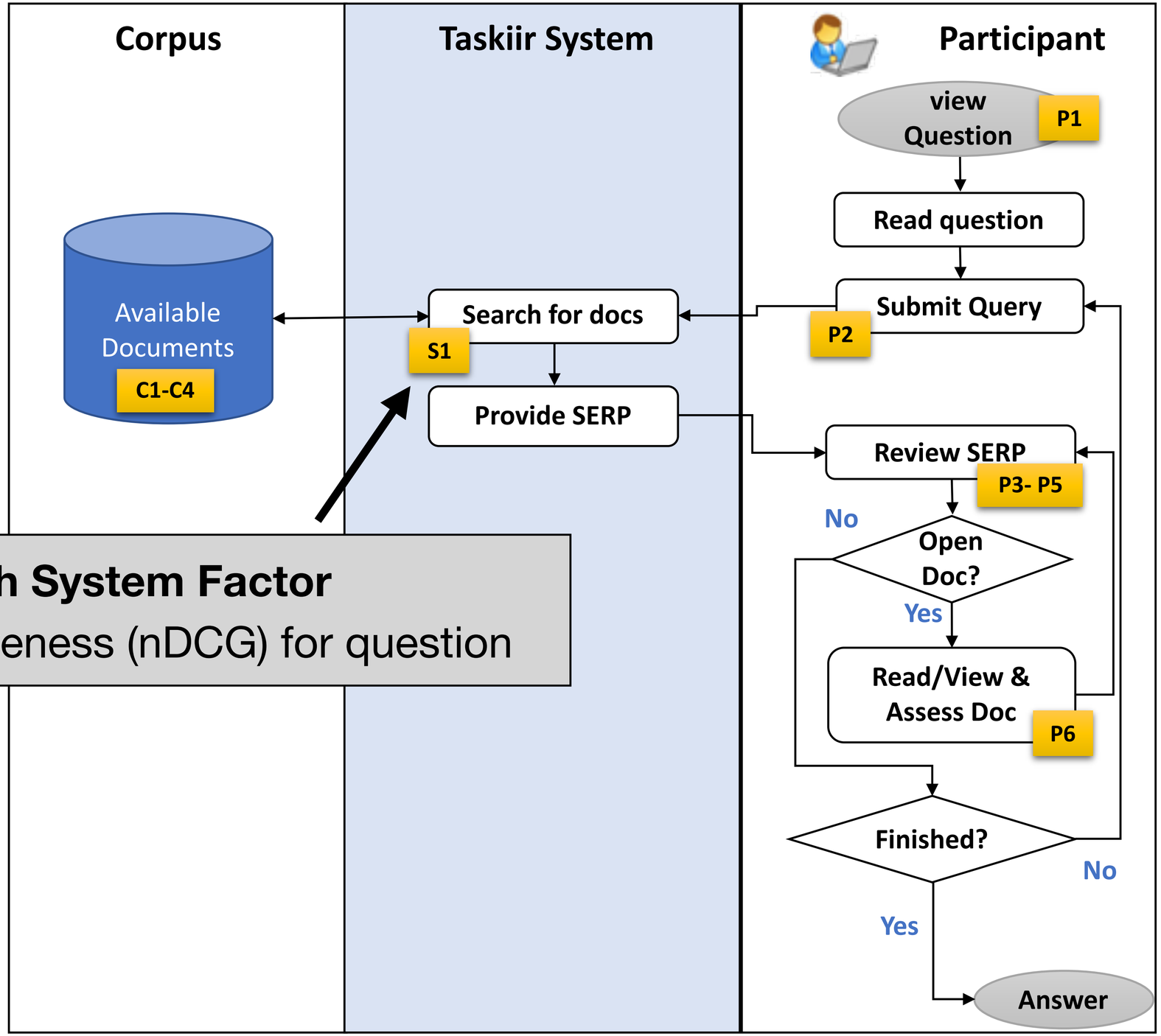
C1: # turning point documents

C2: # derailing documents

C3: ratio turning point / derailing docs

C4: # ambiguous docs

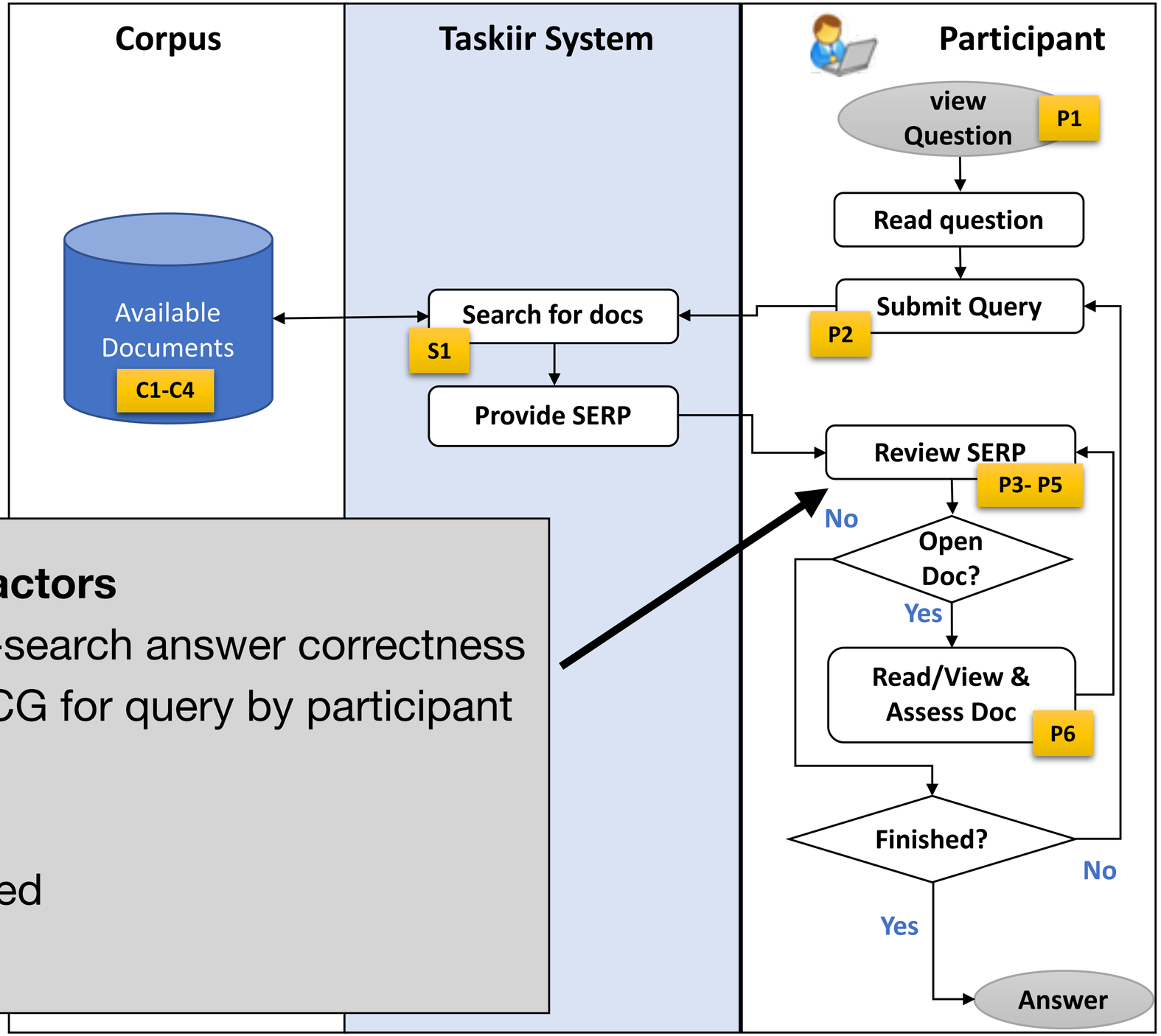
The Success Factors



Search System Factor

S1: search effectiveness (nDCG) for question

The Success Factors



Participant Factors

P1: prior clinical knowledge, i.e. pre-search answer correctness

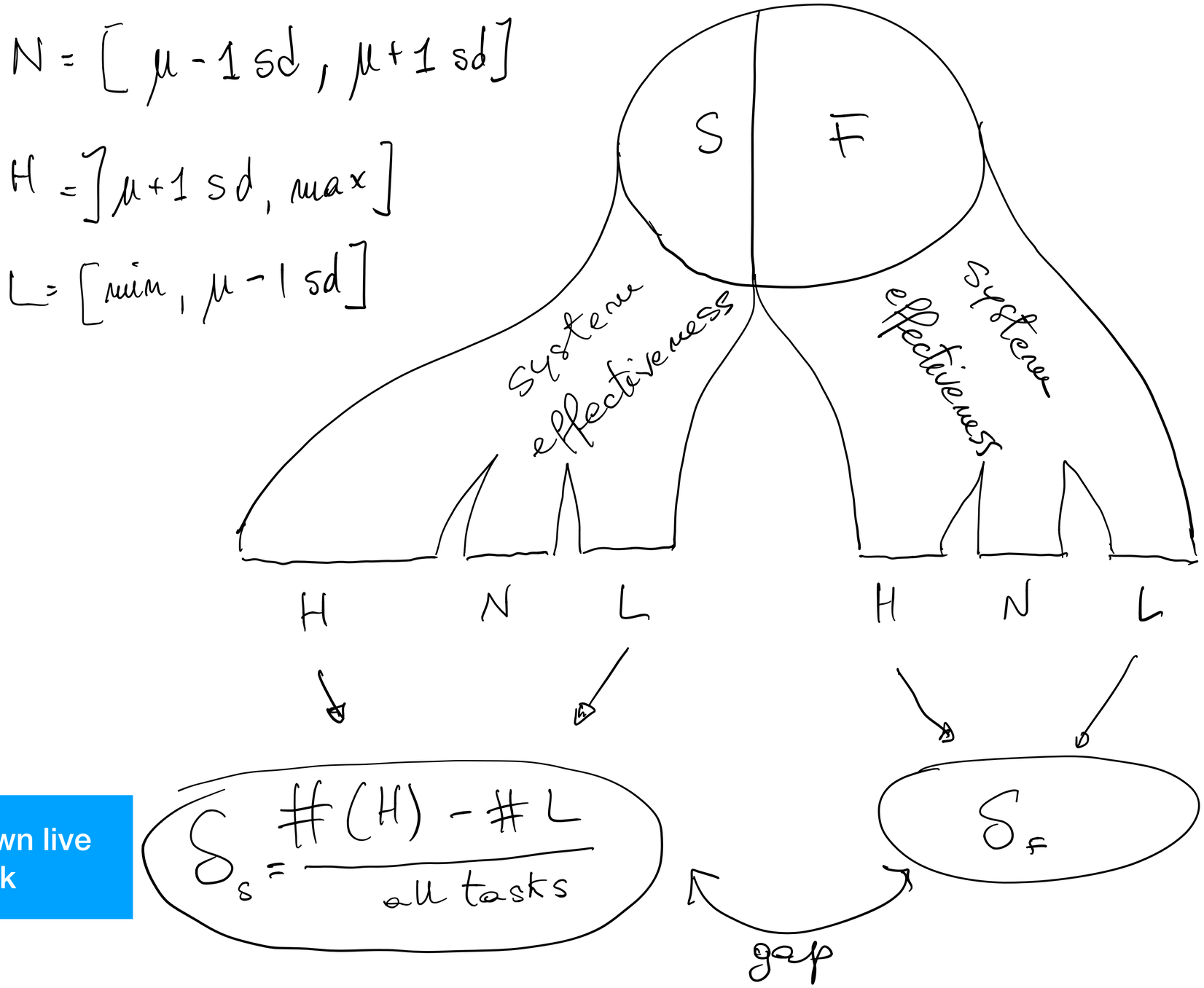
P2: query formulation ability, i.e. nDCG for query by participant

P3: selection turning point docs

P4: selection derailing docs

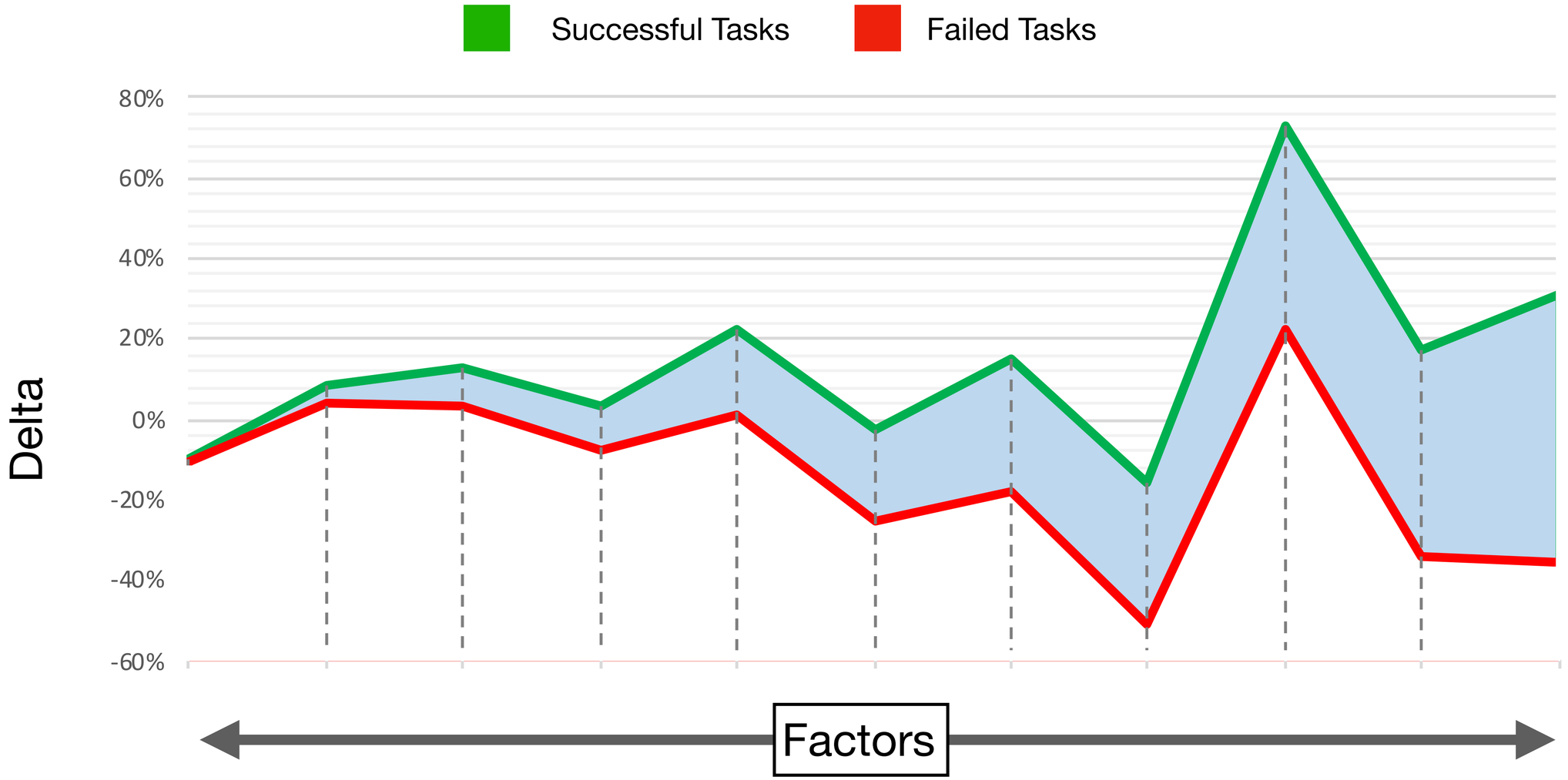
P5: ratio turning/derailing docs viewed

P6: document interpretation

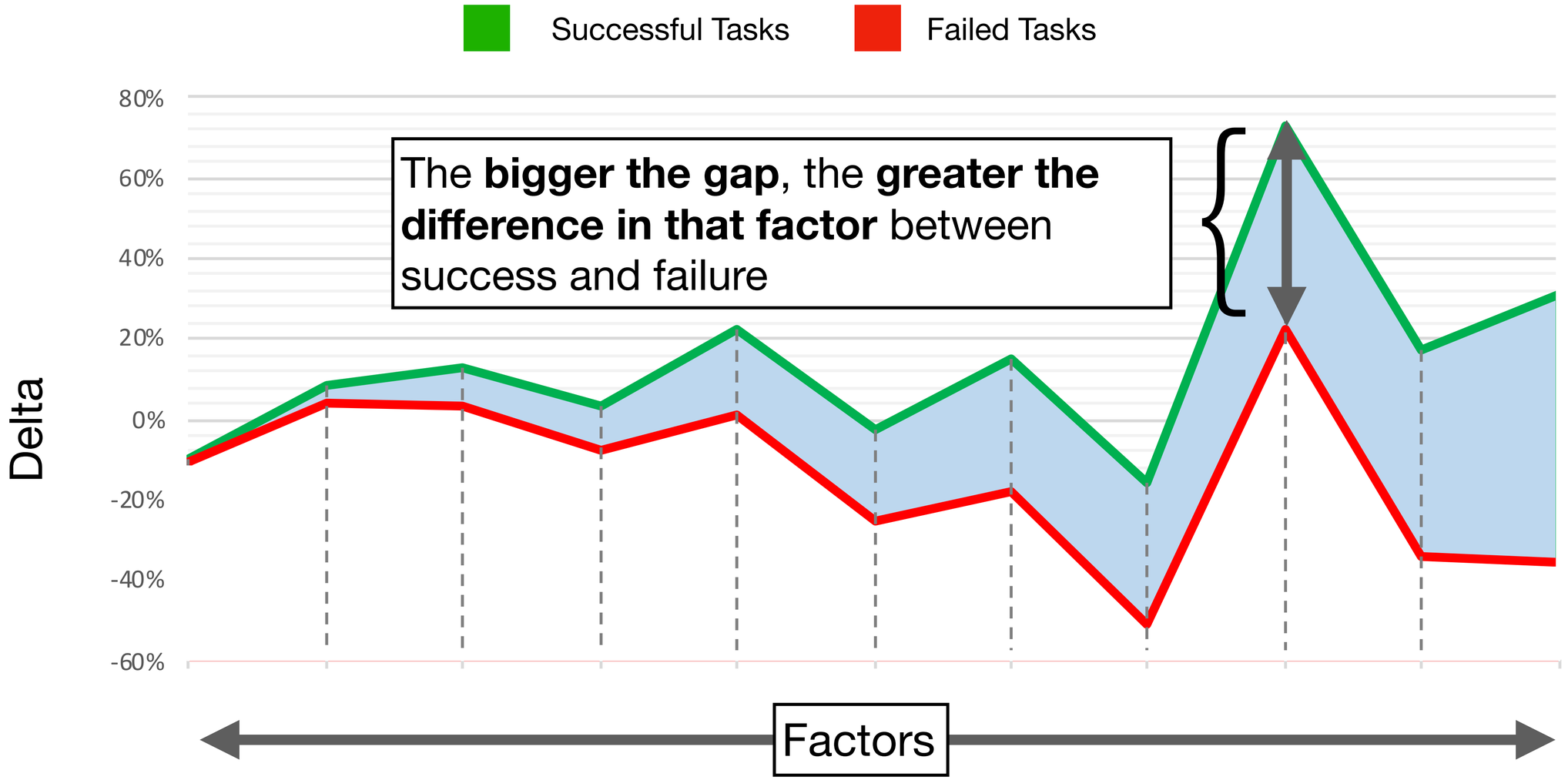


This has been drawn live during the talk

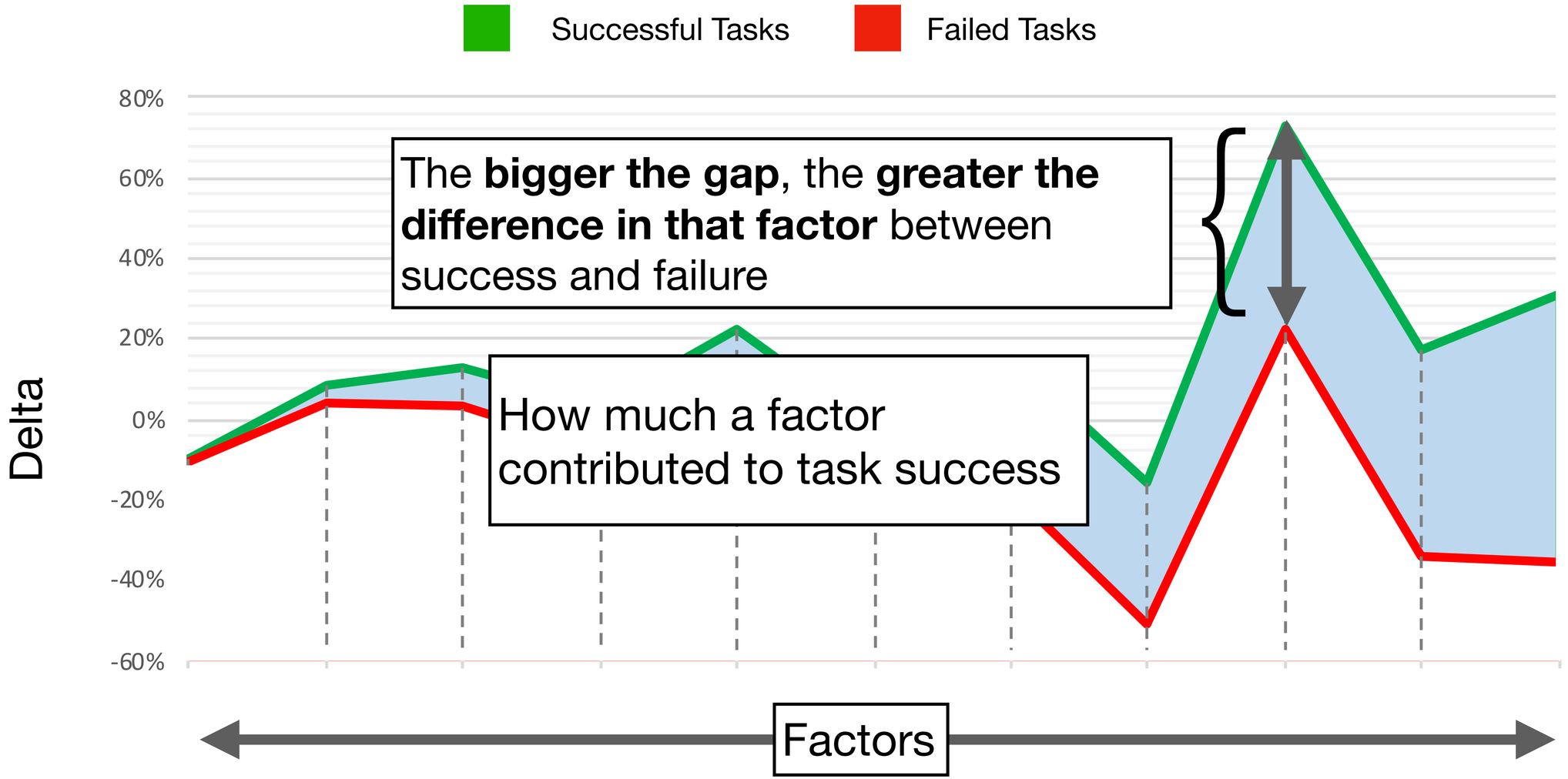
Which Factors are most important?



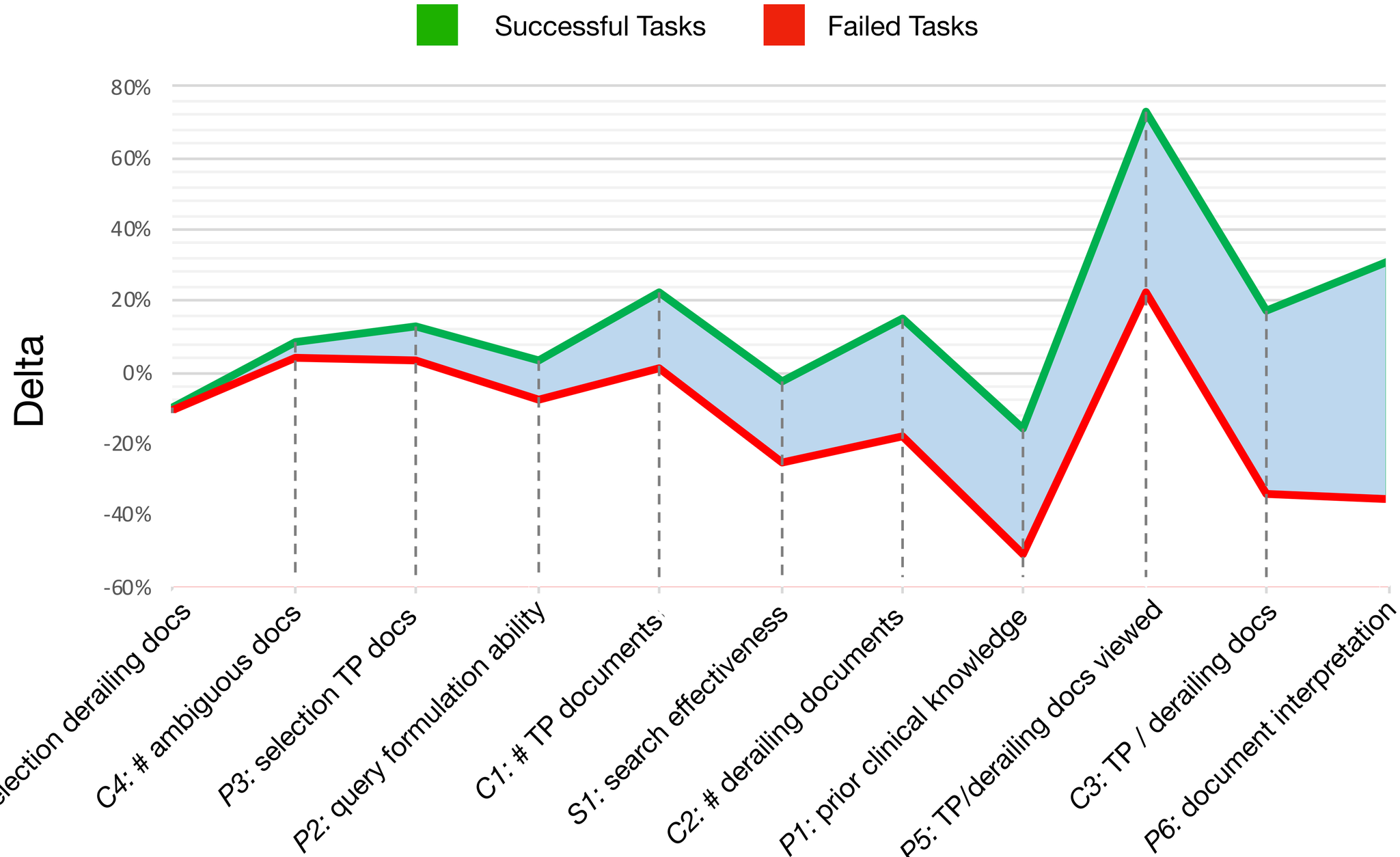
Which Factors are most important?



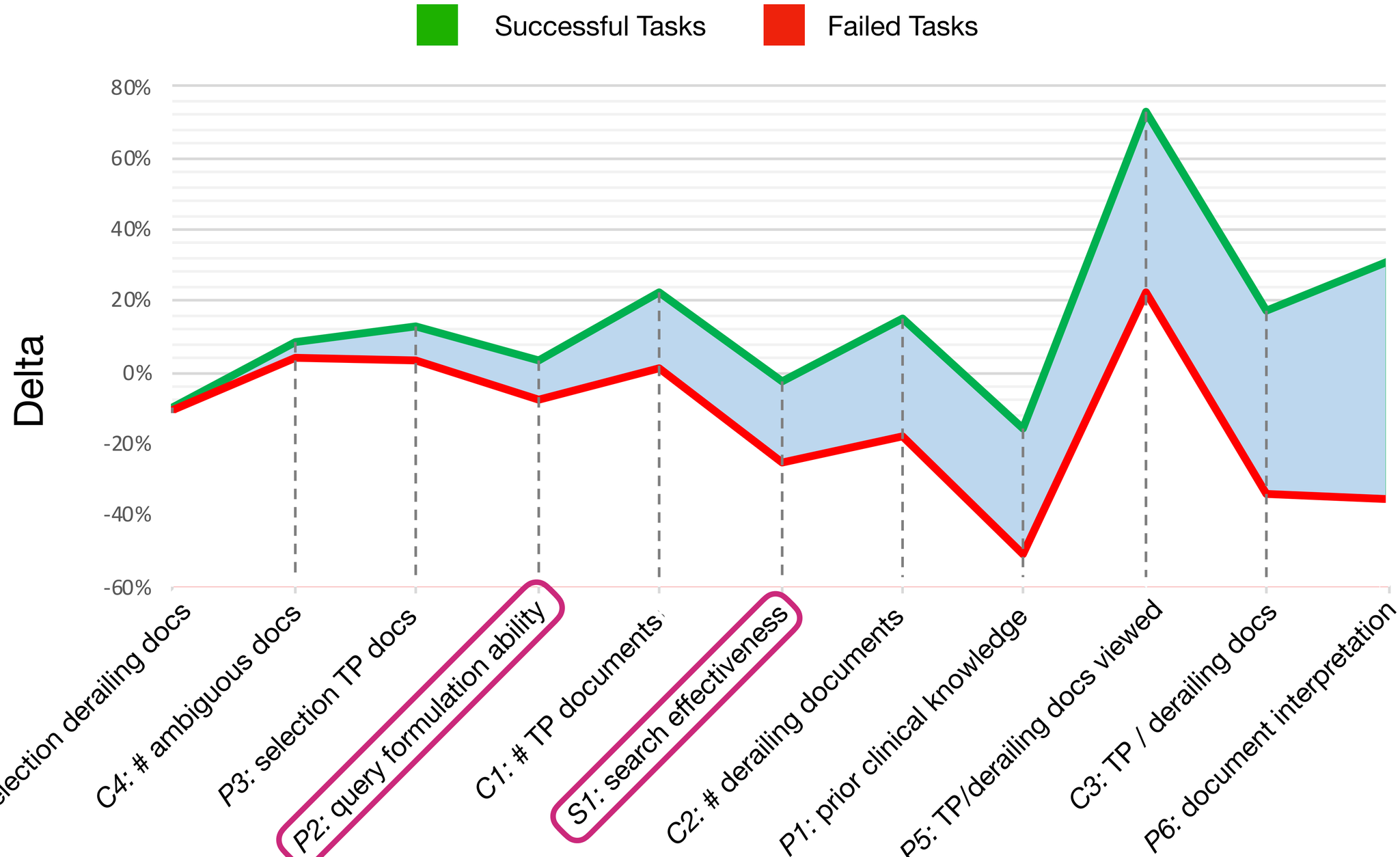
Which Factors are most important?



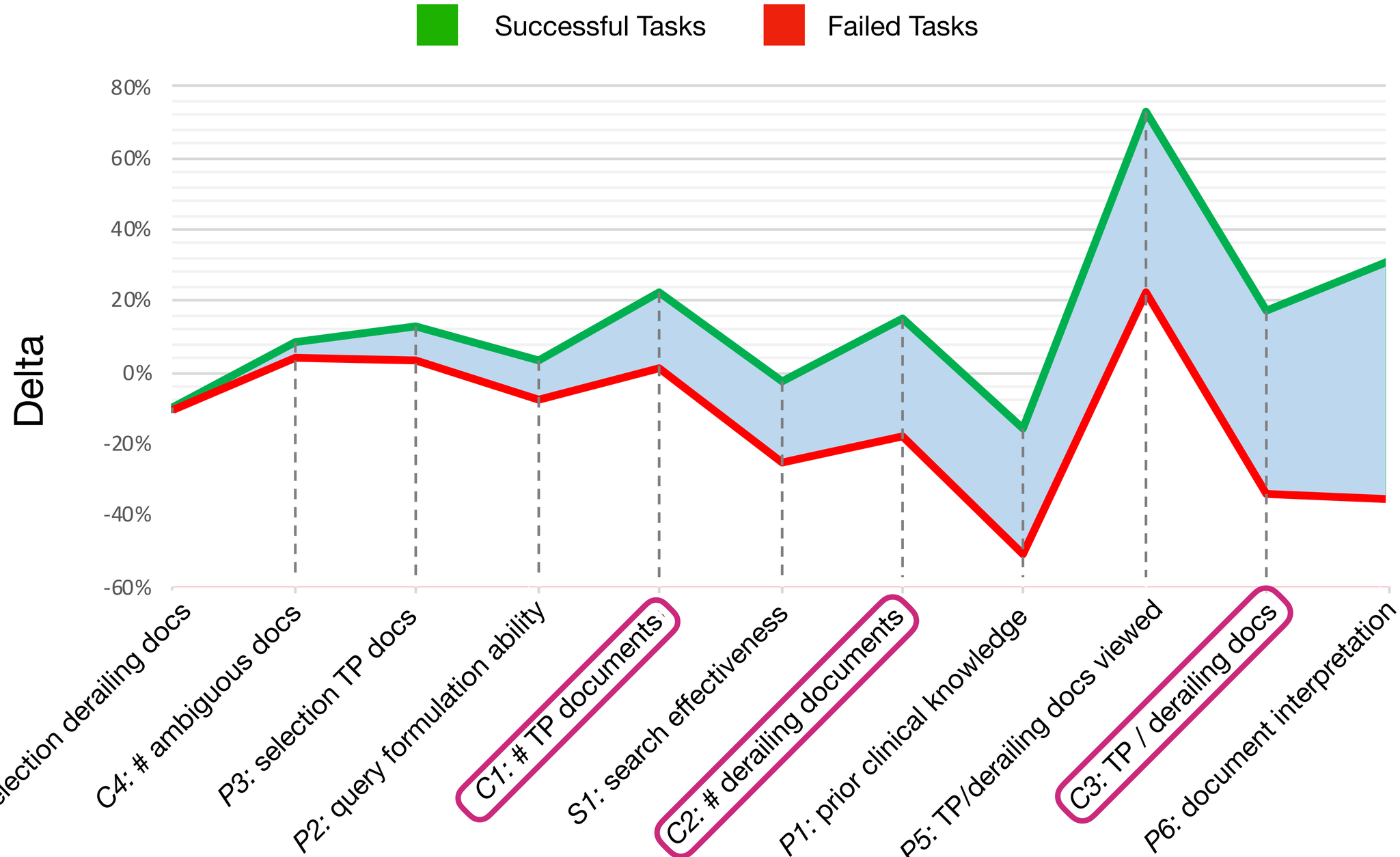
Which Factors are most important?



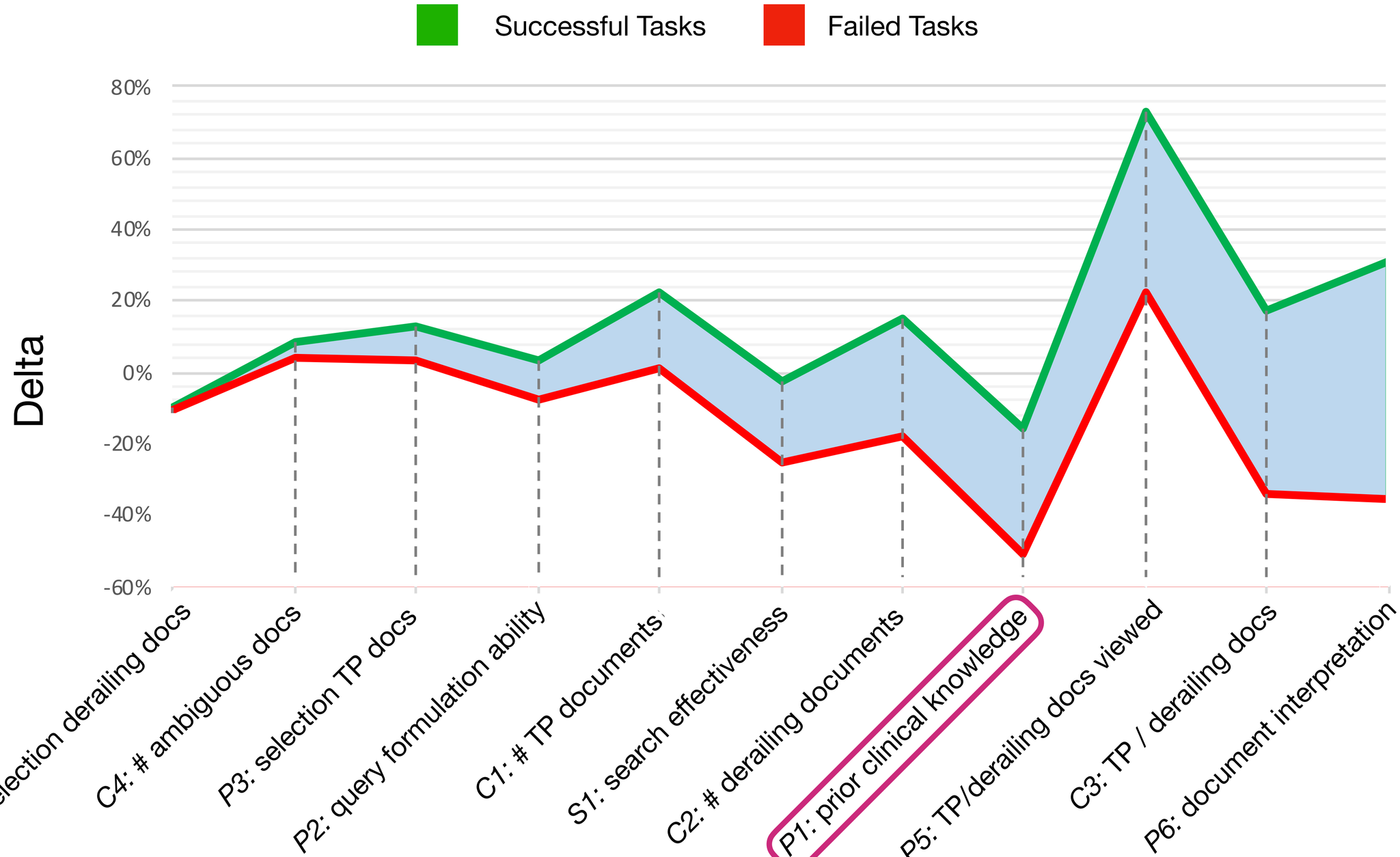
Which Factors are most important?



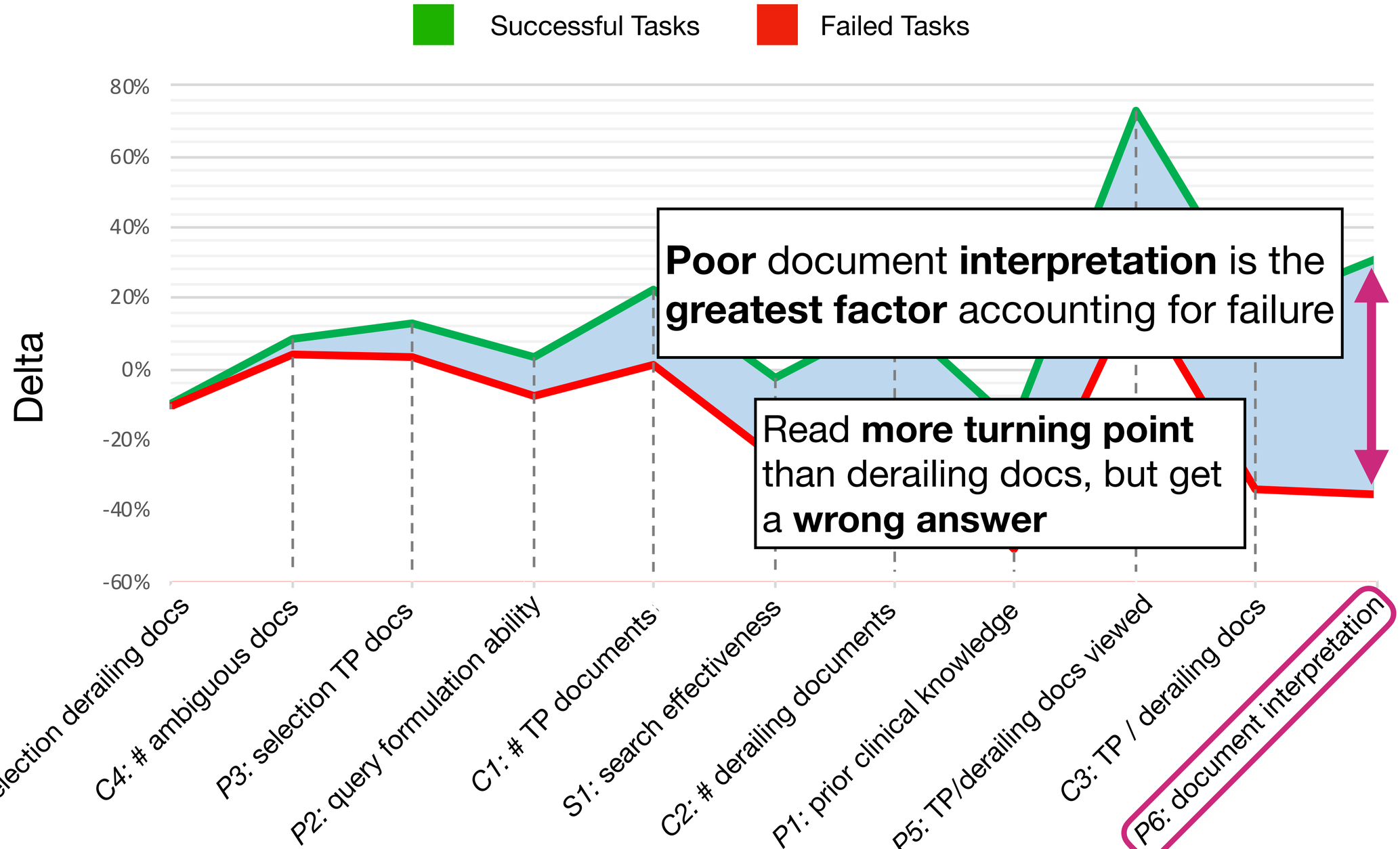
Which Factors are most important?



Which Factors are most important?



Which Factors are most important?



Broader Implications for IR Research

From the results of this study, we have learnt that:

- **System effectiveness** is **not** the only **factor** affecting users decisions
    - actually, it is a weak factor
- Document **misinterpretation, biases** far **more pressing factor** that influence users decisions

- Where would I put my money?

    - **results interpretation**, **resolve** document **ambiguity**, reduce cognitive biases (anchoring, confirmation) [in other domains, also measure correctness, quality of information]
        - How do we **identify derailing** documents?
        - How do we **improve presentation**?
        - How do we help people tie the information to the decision they are trying to make?

- Next study: example of **aiding result interpretation**, in a different, but related task

2nd Study: Consumer Health Search (CHS)

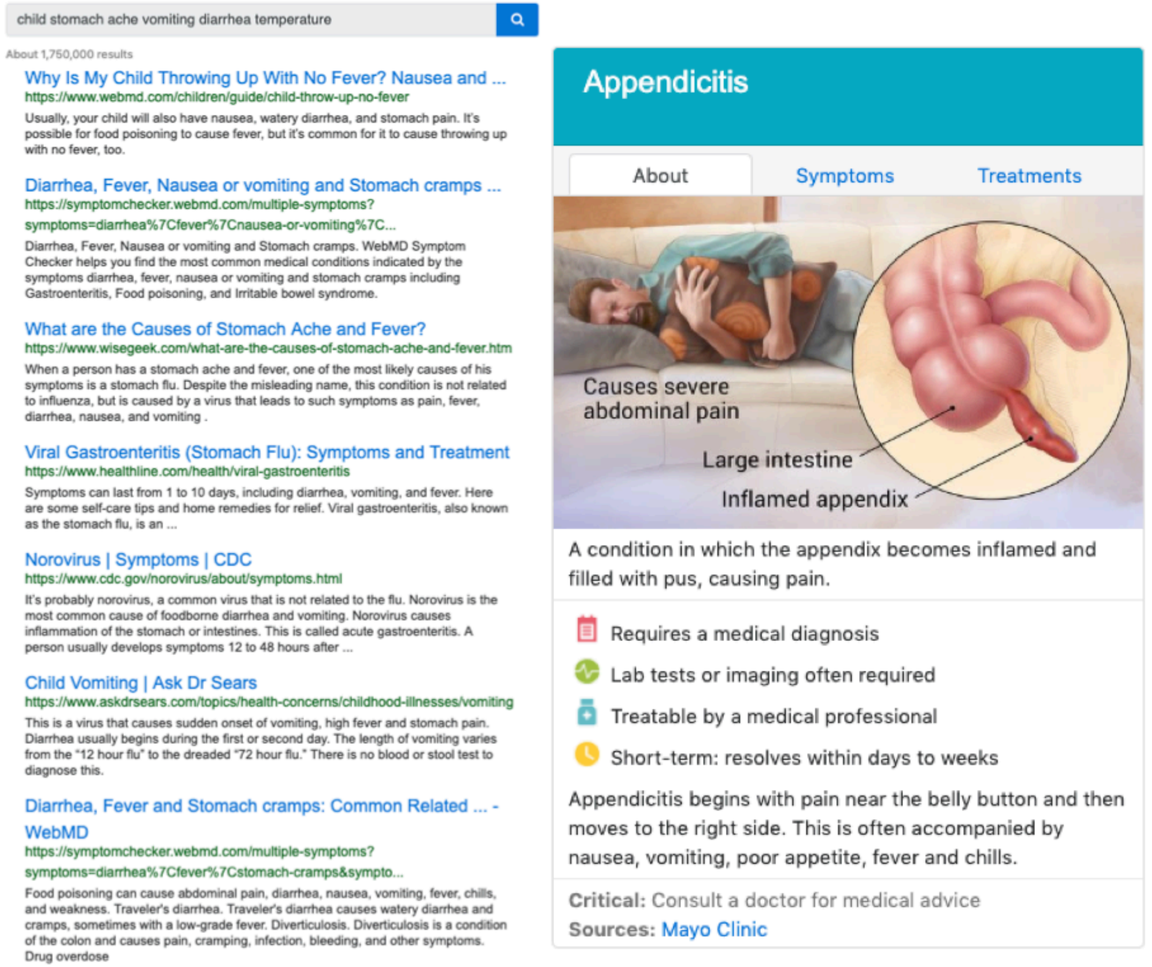
- People search for **health advice** online
  - 59% of U.S. adults has searched online for health information [Fox & Duggan, 2013]
  - 7% of Google searches are health related: ~ 70,000 health queries/minute [Murphy, 2019]
- 59% of self-diagnosers decide **NOT to confirm their condition with a health professional** [Fox & Duggan, 2013]
- Search results strongly **bias people**'s health **decisions** [White, 2009; Dumais et al., 2010; Pogacar, 2017]
- People **struggle to understand** health search results and **fail to make correct health decisions** [Alpay, 2009; Kobayashi & Ishizaki, 2019]
- Search results from **different sources,** with **varying quality**, many web pages **difficult to understand** [D'Alessandro et al., 2001; Becker, 2004]**.** They may be derailing — or even a blocker (dead end)

Assist People Appraising Search Results

- Entity cards used to present **coherent, easy to understand and trustworthy** information (Radlinski, 2015)

- Currently, health cards triggered by queries **that contain health condition name or its aliases**

- We investigate **the benefits of health cards for broader search tasks** - beyond just find out more information about a specific health condition (e.g. symptom-based query)



Based on: (1) Jimmy, G Zuccon, B Koopman, G Demartini, Health Cards for Consumer Health Search, SIGIR 2019
(2) Jimmy, G Zuccon, G Demartini, B Koopman, Health Cards to Assist Decision Making in Consumer Health Search, AMIA 2019
(3) Jimmy, G Zuccon, B Koopman, G Demartini, Health Card Retrieval for Consumer Health Search, CIKM 2019

Do Health Cards Help People Making Effective Health Decisions?

- In SIGIR 2019 we compared search engine with health cards VS no health cards:
  - For a broad range of queries
  - Health cards **reduce effort spent**, **increase good abandonment** rate, and **user's satisfaction**
  - Health cards **benefit** particularly **the not knowledgeable searchers** to identify the correct health information

But, do they make better decisions?

- Next is a study where we focus on **Health Decisions** with respect to **self-diagnosis** (what condition and what to do)

For a health scenario, make the following two decisions:

1. What is the most possible condition for the scenario?

2. What to do / level of urgency ?
   A. Use emergency service
   B. Contact a health professional
   C. Self treat

Sample scenario:

Scenario: Your 12-year-old daughter had a sudden severe abdominal pain with nausea, vomiting, and diarrhea. Her body temperature is 40C.

Correct Diagnosis: Appendicitis

Correct Urgency Level: Emergency

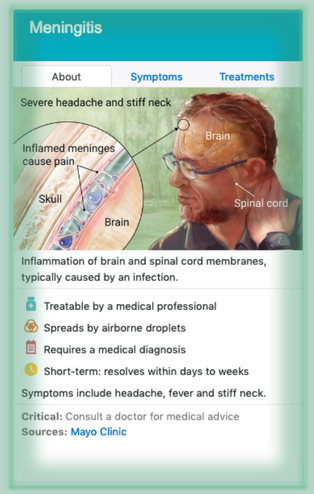
[Semigran et al., 2015]

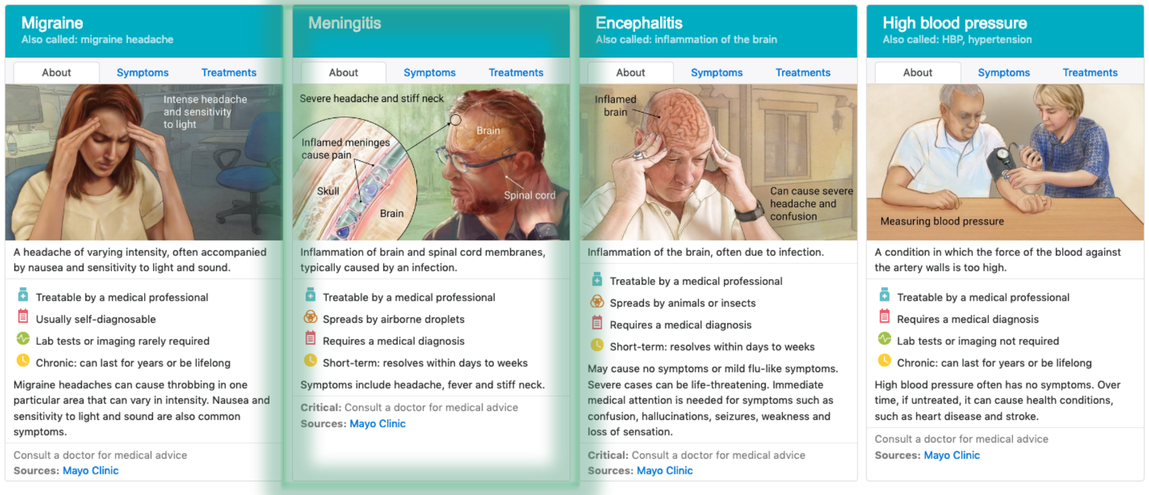
The Four Search Interfaces

Single-Card

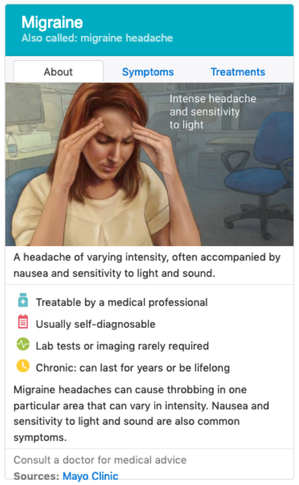
Multi-Cards

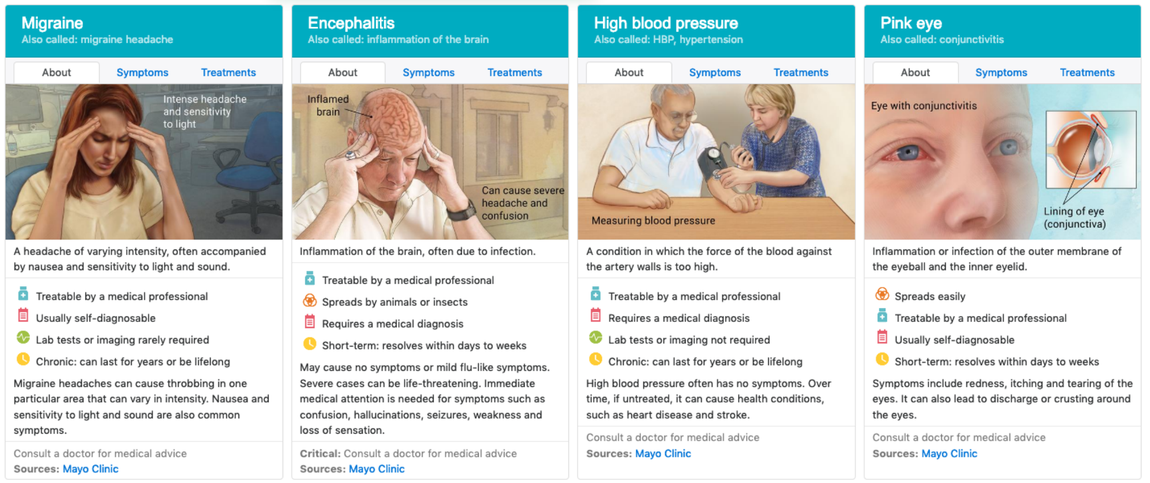
With a Correct Card





No Correct Card



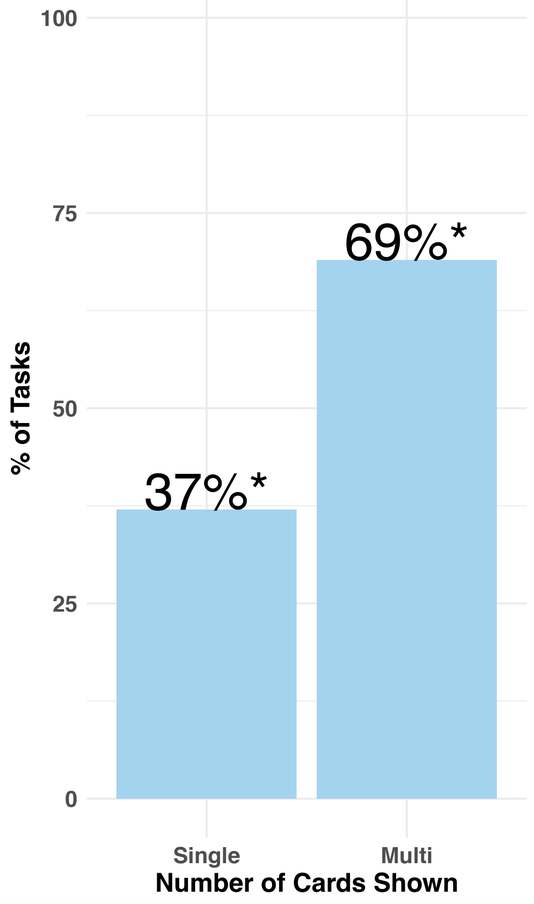


Why Multi-Cards? Presenting possible health conditions side-by-side may enable people to perform **differential diagnosis** + increases chances to present a correct health card

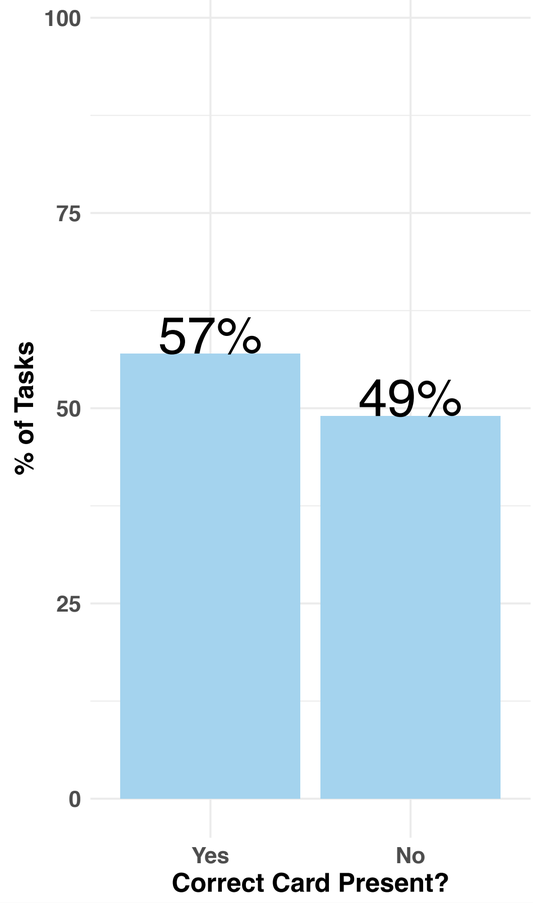
Impact of single and multi-cards?

(1) Health cards selected as a source of information

Information from **multi-cards** selected **significantly more frequently** than from single-card



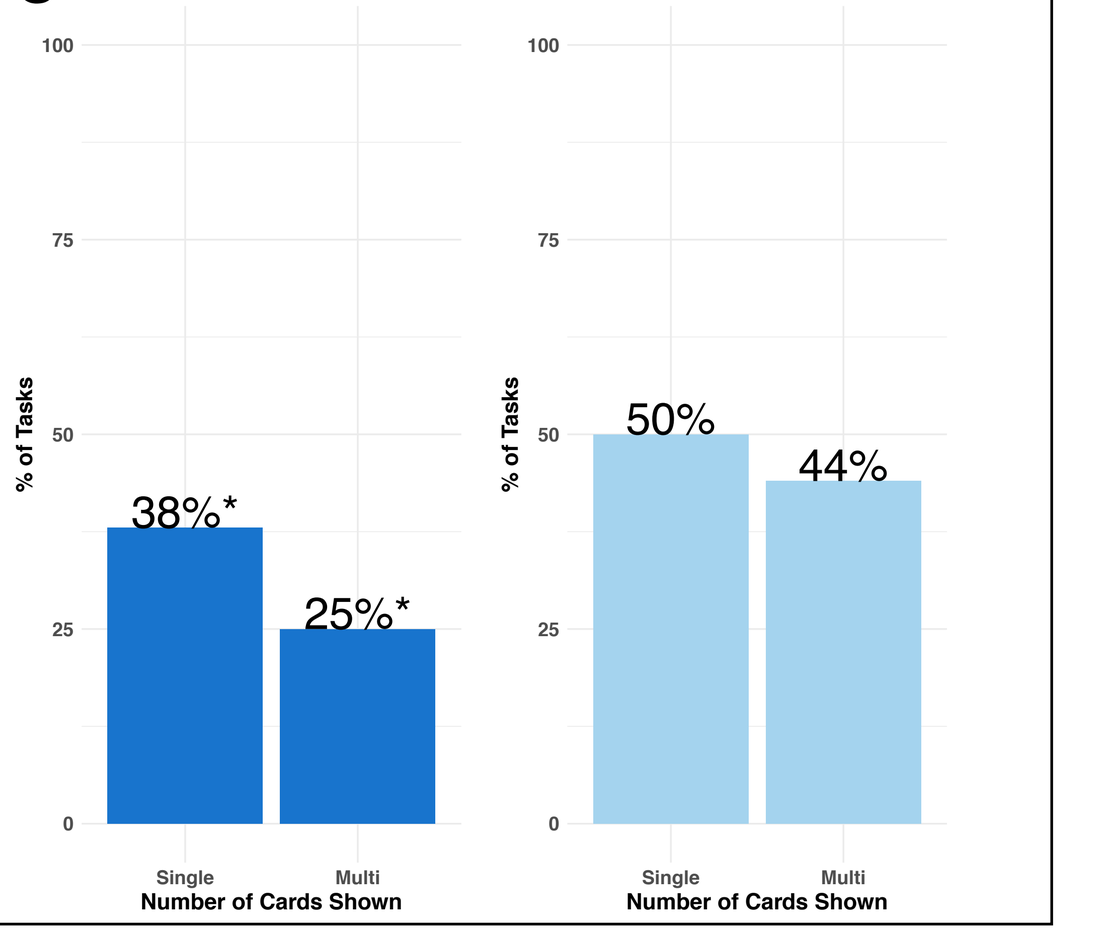
Information from health cards selected **regardless of the correctness** of health card



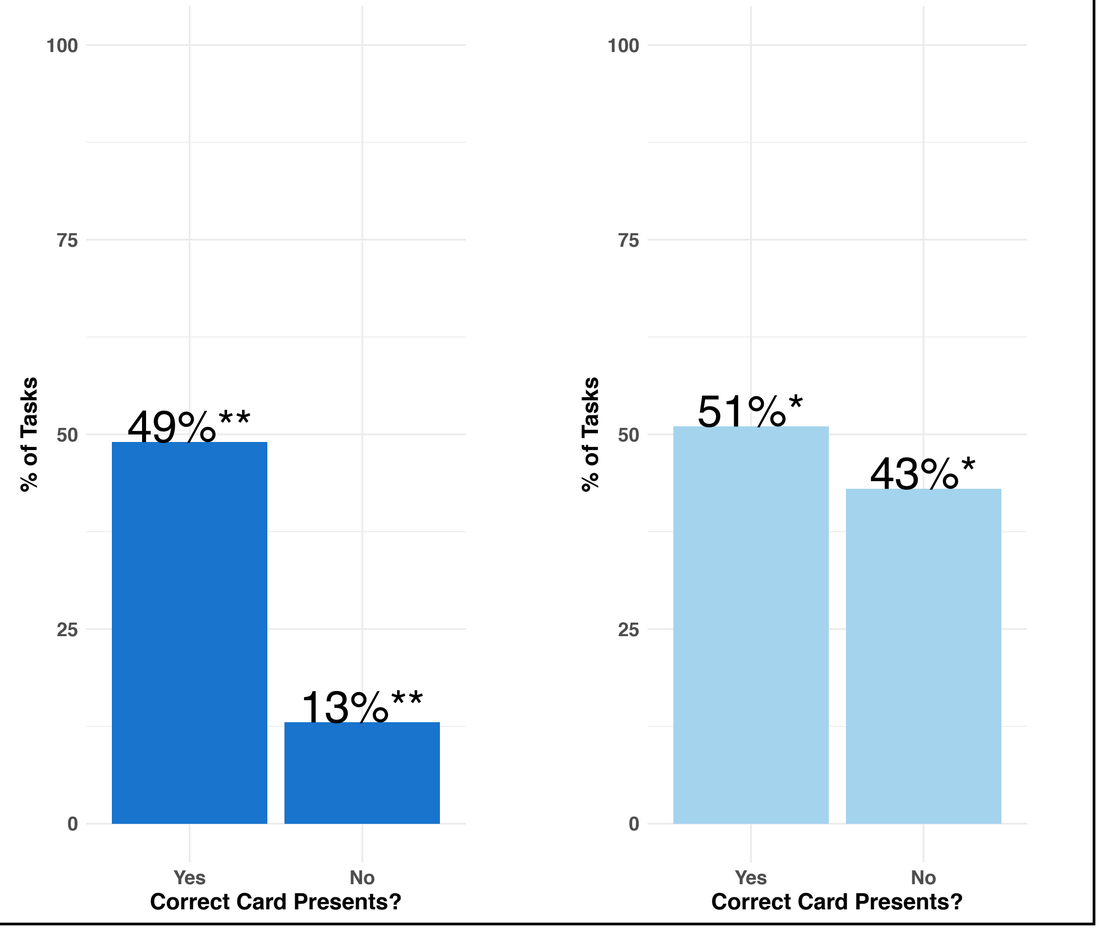
Impact of single and multi-cards?

(2) Correctness of diagnosis and level of urgency

Single card lead to significantly **higher number of correct diagnosis** than multi-cards



Correctness of **decisions** is **dependent** on correctness of **presented health card**



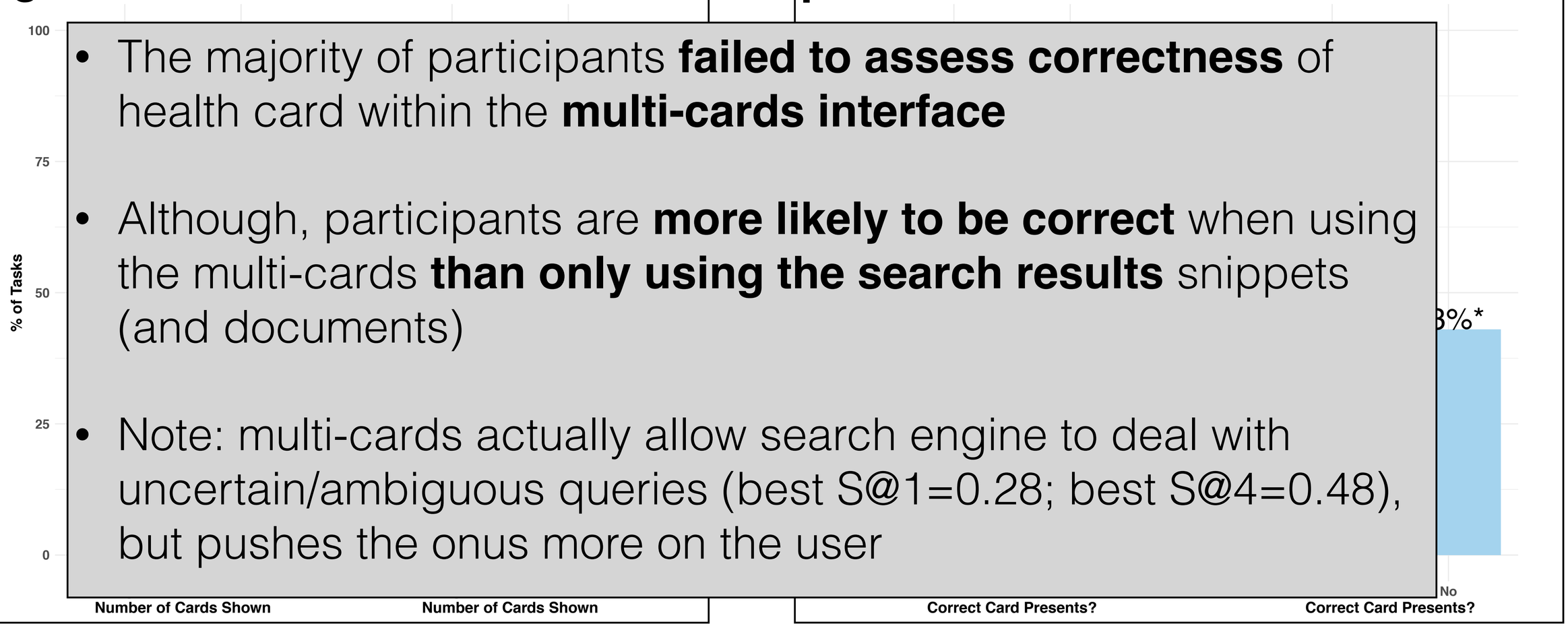
Impact of single and multi-cards?

(2) Correctness of diagnosis and level of urgency

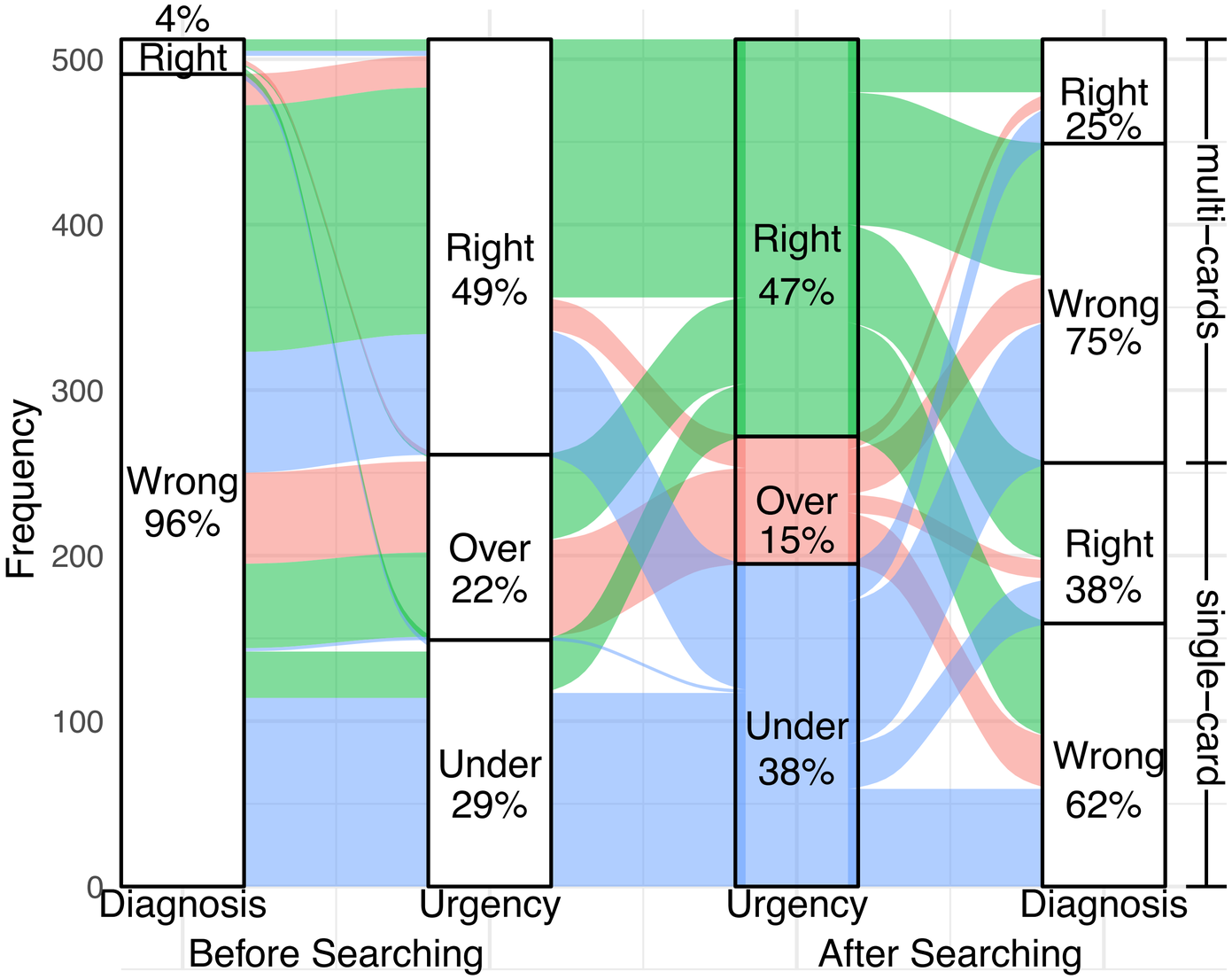
Single card lead to significantly **higher number of correct diagnosis** than multi-cards

Correctness of **decisions** is **dependent** on correctness of **presented health card**

- The majority of participants **failed to assess correctness** of health card within the **multi-cards interface**
- Although, participants are **more likely to be correct** when using the multi-cards **than only using the search results** snippets (and documents)
- Note: multi-cards actually allow search engine to deal with uncertain/ambiguous queries (best S@1=0.28; best S@4=0.48), but pushes the onus more on the user



How does Search Impact Health Decision Making?



- Search activities (with health cards) lead to **significant increase of correct diagnosis**
- Search activities did **NOT** lead to **better** level of **urgency** correctness
- **43%** of **correct diagnosis** had an **incorrect** level of **urgency** (73% of them were **underestimated**)

What have we learnt from this study?

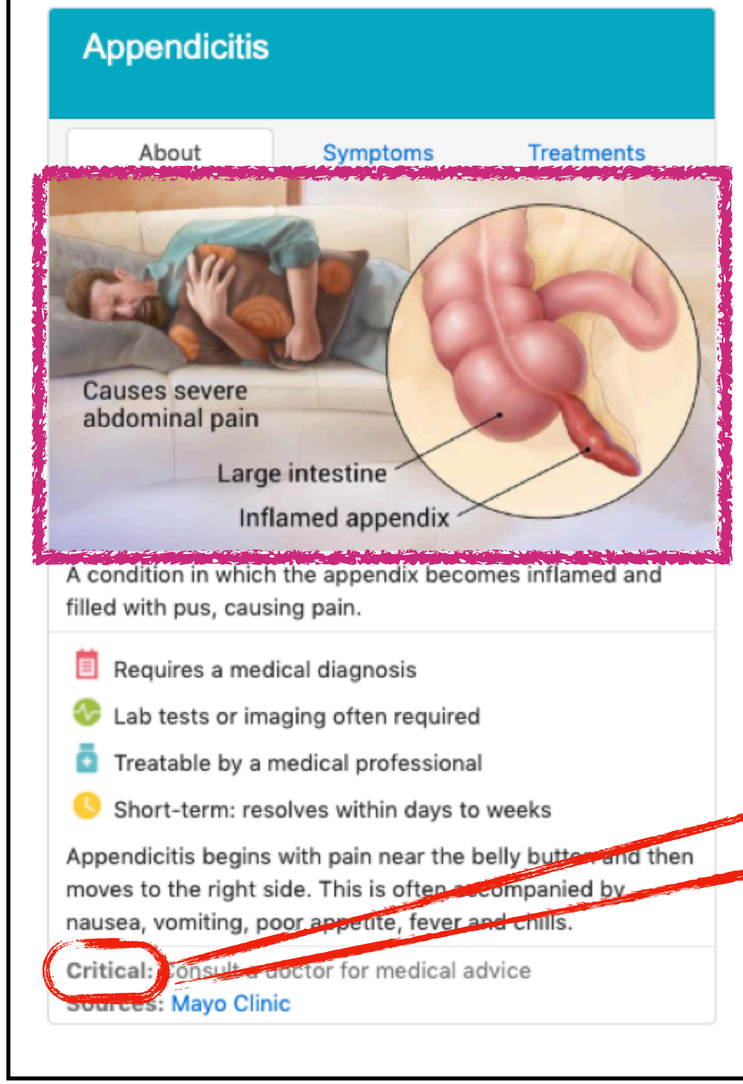
- **Health cards help people** to make more **correct diagnosis**
- **Struggle to infer the level of urgency** — people tend to **underestimate** urgency across search interfaces.
- This is a problem as making the correct diagnosis is not as important as identifying the correct level of urgency

Diagnosis	Urgency Level	Risk
Correct	Under-estimated	Harm well-being + additional health system costs
Correct	Over-estimated	Waste medical resources
Incorrect	Correct	None to minimum

- **Future** work needs to specifically **better assist people in deciding the urgency level** (what to do)

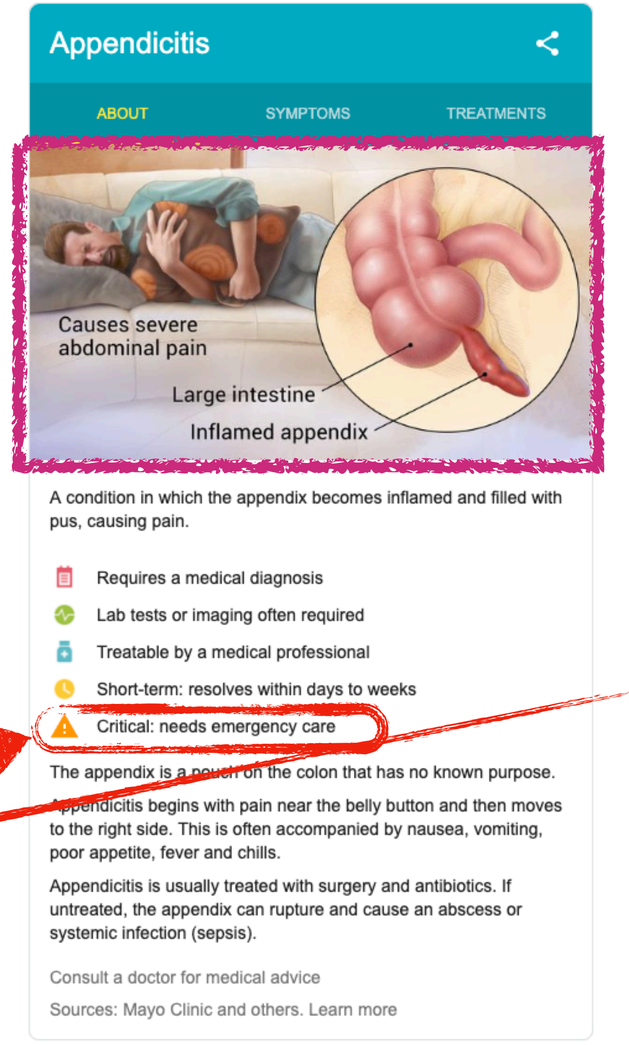
Example of an attempt to address these findings

Google Cards Before Our Study

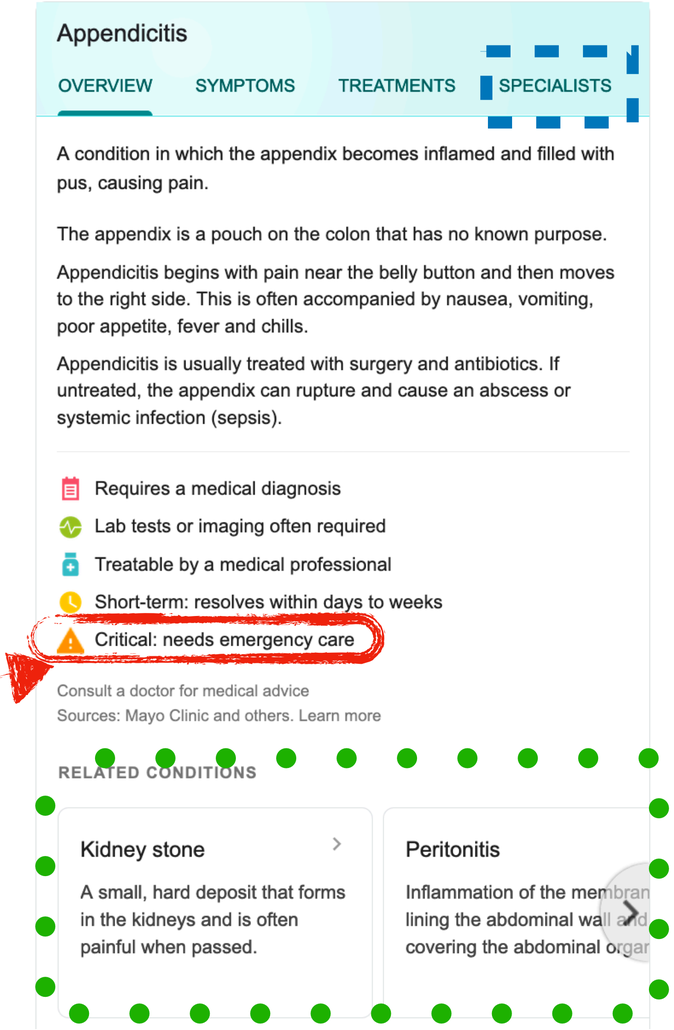


Google Cards Now

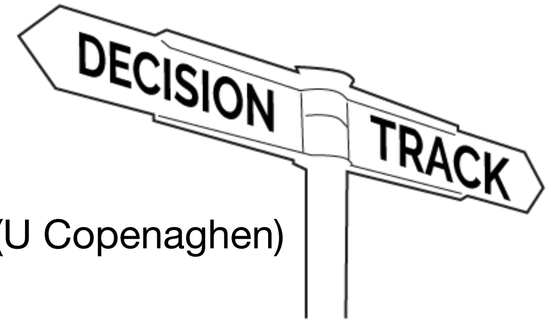
Current Ver. A



Current Ver. B



The TREC Decision Track



Organised with Mark Smucker (U Waterloo), Christina Lioma, Maria Maistro (U Copenaghen)

- Goals of the track:
  - **Foster research** on retrieval methods that **promote better decision making with search engines**
  - Develop **new online and offline evaluation** methods to predict decision quality induced by search results
- It ran first time in 2019 with an ad hoc search task. In 2020:
  - Task 1: (Offline, ad hoc) devise search technologies that **promote correct information** over incorrect information, with the assumption that correct information can better lead people to make correct decisions
  - Task 2: (Online) Given a query, a document ranking (results list) and interaction/behavioural data, **predict** the **decisions** users will take at the end of the search process
  - Task 3: (Offline) Given a query, a document ranking (results list), and assessments (for relevance and correctness), **predict** the **decisions** users will take at the end of the search process
- The task provides assessments of relevance, credibility, correctness + decisions and interactions

Take Aways

- Evaluation drives improvement of search technology. But are we measuring the right thing?
    - **relevance VS decisions**
- It’s not what you think it is relevant, it’s how it is being interpreted
- What do we need?
    - develop a better **understanding of derailing / turning point** documents
    - what are the **blockers** to better decisions?
    - when people have reached a **dead end**?
    - **technology** to support people making **better decisions**

Questions?

g.zuccon@uq.edu.au — ielab.io/guido