Realism in empirical Interactive Information Retrieval studies: A systematic review protocol

Adapted from Review Protocol Template by Sarah Visintini provided by the UNC Health Sciences Library (linked from https://guides.lib.unc.edu/systematic-reviews)

1. Project Information

Review Title	Realism in Empirical Interactive Information Retrieval Studies
Project Lead	Anita Crescenzi, MSIS, PhD
Team Members	Rebecca Carlson, MLS, AHIP Lan Li, MSIS Yu Lee An, MLS, MS(Res), PhD (June-July 2022)
Date	July 2022
Institution(s)	University of North Carolina at Chapel Hill

2. Background

Interactive information retrieval (IIR) studies investigate user's interactions with search systems. As Kelly (2009) notes: "IIR focuses on focuses on users' behaviors and experiences—including physical, cognitive and affective — and the interactions that occur between users and systems, and users and information… IIR evaluation asks the question, can people use this system to retrieve relevant documents?" (p. 2-3)

IIR studies often compare users's performance across multiple search systems or may seek to understand how differences in contextual or situational factors impact a user's cognitive processes or search behaviors when interacting with search systems. In order to make these comparisons, researchers often assign tasks for study participants to complete.

Simulated Work Tasks (SWT)

Borlund (2003, 2016) proposed a method to create "simulated work tasks" (SWT) in to use in IIR studies that describes a simulated information need embedded within a broader (work) task to trigger searching and to provide a reference for the user to assess the relevance of search results. The use of SWTs enables researchers to "simulate" a searcher's genuine information needs.

Using the same (set of) SWT(s) enables researchers to compare search-related outcomes across study conditions through experimental control of the task/information need motivating the search. As Borlund (2016) notes: "The issue of realism of the descriptions of the simulated work task situations is essential

in order for the prompted search behaviour and relevance assessments of the test participants to be as genuine as intended" (p. 396)

Borlund (2016) offers a brief summary of her recommendations for SWTs:

> (1) To tailor the simulated work task situation to the test participants:
> - a situation the test participants can relate to and identify themselves with;
> - a situation the test participants find topically interesting and/or of relevance to them; and
> - a situation that provides enough context in order for the test participants to be able to apply the situation.
>
> (2) To include test participants' personal information needs as baseline.
> (3) To rotate the order of simulated work task situation and personal information needs (counterbalancing).
> (4) To pilot test prior to actual testing (often more than once).
> (5) To display the used simulated work tasks situations when reporting the study. (Borlund, 2016, p. 396)

Borlund recommends tailoring the SWT to the study population in order to provide realism. In addition, asking study participants to search for information to meet their own, genuine information need enables researchers to compare performance across SWT and genuine information needs.

In her 2016 analysis of 67 papers published 1998-2008 using SWTs, Borlund describes how SWTs have been used in studies and the extent to which the studies met the criteria for using SWT. Specifically, she describes the types of evaluations for which SWTs are used, (2) how the SWTs were tailored, (3) whether personal information needs were included as a baseline, (4) whether SWTs were rotated, and (5) whether pilot testing to tune SWTs was reported. Of particular interest for this study, Borlund found a wide range in the quality of tailoring of SWTs with only 3 studies tailoring providing SWTs at the appropriate level and publishing the text of the SWT (p. 403). This resulted in the addition of recommendation #5 above. Borlund also found that only two studies (3%) used personal information needs and none included personal information needs as a baseline (p. 403). In addition, none of the studies described how they pilot tested or tuned their SWTs (p. 404).

The recommendations by Borlund (2003, 2016) for creating and using SWTs have been widely used and have helped researchers with some of the many experimental design decisions needed to conduct experimental IIR studies with users. In our experience, the foundational work and recommendations made by Borlund recommendations have been critical to our design and interpretation of experimental IIR studies.

As with the studies in the meta-analysis, we have implemented an subset of Borlund's recommendations. In Crescenzi (2019), we followed recommendations 1, 4, and 5: we carefully tailored the SWTs to our study population, we assessed the SWTs in pilot testing and in a separate study, and we report an example SWT in Crescenzi and Li (2022) and the full text of all SWTs in Crescenzi (2019). We did not follow recommendations 2 and 3: using personal information needs as a baseline and counterbalancing SWTs and personal information needs. Although we recognize the potential value of being able to compare genuine and simulated information needs (i.e., participant-generated tasks vs.

researcher-imposed SWTs), we were concerned that following recommendations 2 and 3 would introduce an apples and oranges comparison. To what extent can which performance and perceptions are comparable for searches triggered by a personal information need vs. a simulated information need (or set of simulated information needs)? In a controlled experimental study, a researcher seeks to identify and control for potentially confounding factors through experimental design or statistical control. Although using SWTs enables researchers to have more experimental control, using participant-generated information needs introduces potential confounds. What if the user brings an information need that does not match the parameters set by the researcher? What if the scope of the personal information need and SWT are different? In addition, we had questions about the extent to which a "real-life information need" that a participant brings into a study after being requested to do so by a researcher truly represent their actual need for information. To what extent will it have been modified to meet parameters requested by the researcher? Even if the information need represents a real life information need, how do other factors relating to the study design modify the "realism" of the participant's own personal information need (e.g., study timing, contents of search system)?

By not recommendations 2 and 3, we wondered how we could ascertain whether we created SWTs that would trigger realistic cognitive processes and be reflected in measures of search behavioral traces. As Borlund (2016) notes, "the personal information needs become the tool to compare, interpret, and validate the test participants' interaction patterns achieved by use of the simulated work task situations." (p. 403)

Realism in experimental studies

In experimental research, realism can be present on multiple levels: mundane, experimental, and psychological. Wilson, Aronson, and Carlsmith (2010) describe two types of realism first described in Aronson and Carlsmith (1968):

> "In one sense, an experiment is realistic if the situation is involving to the participants, if they are forced to take it seriously, if it has impact on them. This kind of realism they called *experimental realism*. In another sense, the term "realism" can refer to the extent to which events occurring in the research setting are likely to occur in the normal course of the participants ' lives, that is, in the "real world. " They called this type of realism *mundane realism*" (p. 54)

Wilson et al. (2010) also describe psychological realism first described in Aronson, Wilson, and Akert (1994):

> This is the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life. It may be that an experiment is nothing like what people encounter in everyday life (low in mundane realism) and fails to have much of an impact on people (low in experimental realism). It could still be high in psychological realism, however, if the psychological processes that occur are similar those that occur in everyday life.(p. 55)

As we noted in Crescenzi and Li (2022).

> Multiple aspects of realism are important for experimental researchers to consider... mundane, experimental, and psychological realism. These aspects of realism are important in IIR studies: if

> the scenario and tasks presented to the participant are similar to what they might do in their real life (mundane realism), if participants get involved in the tasks and take them seriously (experimental realism), if cognitive processes are similar to those in real life (psychological realism), and if participants' behaviors are similar to those that would be observed in real life. (p. 266).

In contrast to comparing performance between SWTs and personal information needs as recommended by Borlund, some decision-making studies have used "realism check" questions (similar to manipulation check questions) to check whether the researcher-imposed scenarios used in experimental decision-making studies are realistic to study subjects. For example, Dabholkar (1994) had student participants rate the realism of researcher provided scenarios used two questionnaire items ("the situation described was realistic" and "I had no difficulty imagining myself in this situation") on a seven-point Likert scale.

Darley & Lim (1993) go one step further and issue a call to include realism checks in experimental studies:

> Although we would be the first to admit that creating realism and involvement checks is a difficult endeavor, the importance of ensuring experimental realism calls for nothing less. Thus, we propose that every experiment should attempt to incorporate such checks. These checks could include items that measure the perceived meaningfulness and artificiality of the experimental task, the respondents' degree of involvement in the experimental task, and the perceived relevance of the experimental roles. These checks should be performed in the pilot study so that screened information can be used to design a more realistic main-study experiment. In addition, we recommend that these checks be included in the main study to ensure experimental realism. (Darley & Lim, p. 493)

Realism in IIR experimental studies

Although it is common in IIR studies for researchers to use SWTs in empirical studies, it is less uncommon for researchers to discuss or assess the realism of the study or SWTs used. This study seeks to better understand the extent to which researchers discuss realism when they use SWTs and how they discuss realism.

Capra, Velasco-Martin, and Sams (2011) found similar levels of self-reported engagement on researcher-imposed exploratory decision tasks vs. self-generated tasks although participants reported higher engagement with self-generated tasks vs. imposed tasks generally. Crescenzi (2019) adapted the realism check questions from Dabholkar (1994) to assess participant performance on and perceptions of everyday life decision tasks created for the study. These realism check questions were used during a preliminary study that had the goal of assessing the tasks, and in a second study that used the tasks to compare search and decision-making behaviors and perceptions between two experimental conditions. In addition to realism check questions, post-study interviews probed the realism of the scenario ("make recommendations for a friend") and each decision topic experienced during the study (e.g., mesh wifi, short-term housing options). Brief results of the realism assessment are presented in Crescenzi (2019) and more details are presented in Crescenzi and Li (2022).

3. Objective

This systematic review expands upon on the meta-analysis of SWT use in IIR studies (Borlund, 2016) to analyze the extent to which realism in IIR experimental studies is discussed and assessed, and any discussion of a potential impact of realism (or lack of realism) on IIR study outcomes.

Our systematic review aims to identify empirical studies (C4) of people (C1) conducting interactive searches in which they initiate a request for information from one or more information retrieval systems (C2) to complete researcher-assigned tasks (C3). A subset of these studies that explicitly discuss the realism (C5) of elements or aspects of the study (e.g., realism of researcher-assigned tasks, IR systems, etc.) will be analyzed in-depth.

Specifically, we will address several research goals and questions. Our first goal is to identify studies that use SWT and the subset that also discusses realism. Our second goal is to more closely analyze studies that explicitly discuss the realism of one or more aspects of the study design. all of the studies.

RQ1: For papers that use SWT (and talk about realism): a) describe study including the (i) study participants, (ii) assigned tasks, (iii) study design, b) describe how they talk about realism (i) in general, (ii) of assigned tasks, and (iii) of participant-generated tasks.

RQ2: For papers that use SWT and talk about realism, how did they talk about steps they took to ensure realism of the assigned tasks (including tailoring) or the overall study design?

RQ3: For papers that use SWT and talk about realism, how did they talk about how they assessed the realism of the assigned tasks or the overall study design?

RQ4: For papers that use SWT and talk about realism, how did they talk about an impact or potential impact of realism on their results?

4. Search Strategy

Databases

- LISS [Link], https://guides.lib.unc.edu/go.php?c=23609515
- LISTA [Link], https://guides.lib.unc.edu/go.php?c=23608980
- ACM Digital Library [Link], https://guides.lib.unc.edu/go.php?c=23608308
- Scopus [Link], https://guides.lib.unc.edu/go.php?c=23609180
- ASIST DL: [Link], https://asistdl.onlinelibrary.wiley.com/search/advanced?text1=

Hand searching

A comprehensive list of publication venues for IIR studies was compiled based on previous systematic reviews (e.g., Kelly & Sugimoto, 2013) and the research team's expertise. This list of publication venues was compared to the coverage of the databases listed above in June and July 2022. The full list of IIR

publication venues, their years of coverage in the databases, and the years/issues that are not included in the database can be found in the accompanying Appendix.

Publications and years not available in the databases to hand search (as of 7/18/2022) include

- Annual Meeting of the Association for Information Science and Technology
  - 1997, 2000-2001
- European Conference on Information Retrieval
  - Pre-2005
- European Conference on Research and Advanced Technology for Digital Libraries
  - Pre-2005
  - Note: In 2011, renamed to TPDL (International Conference on Theory and Practice of Digital Libraries).
- Symposium on Human-Computer Interaction and Information Retrieval (HCIR)
  - Indexed in ACM DL 2012-2013
  - https://sites.google.com/site/hcirworkshop/
- iConference
  - Pre-2011, 2013-2017
- Information Seeking in Context
  - Note: Published as a supplement to the Information Research journal but not indexed in Scopus with Information Research (or any other journals).

We will also search the Repository of Assigned Search Tasks (RePAST) for any mentions of concept 5 (i.e., realism). RePAST is available electronically and has been described in several publications (Wildemuth & Freund, 2009, 2012).

- *https://ils.unc.edu/searchtasks/index.html*

Experts/stakeholders

Five experts in Interactive Information Retrieval were asked in June 2022 if they would review the list of databases and IIR publication venues to ensure coverage of relevant publications. Our goal is to have outside vetting of our list by at least one expert.

Reference searches

We will perform citation tracing of seven key publications authored by Borlund:

1. Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, *8*(3), 8-3. http://informationr.net/ir/8-3/paper152.html
2. Borlund, P. (2016). A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *Journal of Documentation*. *72*(3), 394-413. https://doi.org/10.1108/JD-06-2015-0068 [Link]
   a. 67 papers using simulated work tasks were included in this meta-evaluation.

3. Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of documentation, 53*(3), 225-50. https://doi.org/10.1108/EUM0000000007198
4. Borlund, P., & Ingwersen, P. (1998, August). Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 324-331). ACM. https://doi.org/10.1145/290941.291019
5. Borlund, P., & Ingwersen, P. (1999). The application of work tasks in connection with the evaluation of interactive information retrieval systems: empirical results. *MIRA'99: Proceedings of the 1999 International Conference on Final Mira* (pp.1-15). BCS Learning & Development. https://dl.acm.org/doi/10.5555/2228065.2228066
6. Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation, 56*(1), 71-90. https://doi.org/10.1108/EUM0000000007110
7. Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, *54*(10), 913-925. https://doi.org/10.1002/asi.10286

5. Eligibility Criteria

Our systematic review aims to identify empirical studies (C4) of people (C1) conducting interactive searches in which they initiate a request for information from one or more information retrieval systems (C2) to complete researcher-assigned tasks (C3). A subset of these studies that explicitly discuss the realism (C5) of elements or aspects of the study (e.g., realism of researcher-assigned tasks, IR systems, etc.) will be analyzed in-depth.

Our systematic search strategy maps to five concepts: people/users (C1), interactive information retrieval (C2), assigned tasks (C3), empirical studies (C4), realism (C5). Table 1 includes our concepts, related inclusion and exclusion criteria, and sample query syntax. These are mapped to the SPIDER framework that we used to create and refine our concepts, eligibility criteria, and data extraction elements. SPIDER was adapted from PICO and designed for systematic reviews of non-quantitative data, i.e., for qualitative and mixed methods research (Cooke, Smith, & Booth, 2012).

In addition, we limit our search to peer-reviewed publications, written in English, and published 1997 and 2022. In 1997, the first of Borlund's papers discussing realism and simulated work tasks was published. These criteria will be used to screen all papers identified through database searching, hand searching, and citation chasing.

Two rounds of database searches will be completed. One that searches for the intersection of concepts 1-4, and one that searches for the intersection of concepts 1-5. For succinctness, these will be referred to as the "ST" search or set of papers and "realism" search or set of papers.

Table 1: Systematic review concepts, eligibility criteria, and query mapped to the SPIDER framework.

Concepts	**Eligibility criteria for search and screening (* criteria used in Kelly & Sugimoto, 2013)**	**SPIDER *[PICO]***	**Draft query** (v8, 6/17/2022)
Users (C1)	Inclusion • Humans must be included as test subjects.* • Crowd workers who meet other criteria are included.	Sample *[Population]*	*"user" OR "users" OR "human" OR "humans" OR "subject" OR "subjects" OR "participant" OR "participants"*
Interactive Information Retrieval (C2)	Inclusion • The purpose of the study should be for people to use a search system to accomplish some task(s). This includes but is not limited to evaluations of an IIR system or feature.* • Subjects must engage in information retrieval with interactive searching, where they initiate the search (e.g., enter a query) and evaluate results. * • Studies in which subjects engage in cognitive processing of the search results (e.g., assess the relevance of document) are included. • Studies in which researchers give queries to participants to use with the search system are included. • Search or information retrieval systems are not limited by the information that they search (e.g., open web, news corpus, dataset). • Search systems are not limited to text-based search systems only. Multi-media and other non-text-based search systems are included. Exclusion • Studies in which a system only pushes information to users without users requesting it (e.g., filtering and recommender systems) are excluded. • Annotation studies are excluded as the goal is to collect relevance assessments (e.g., to use as ground truth relevance in system-focused evaluations of information retrieval systems).	Phenomenon of Interest *[Intervention]*	*"interactive information retrieval" OR "IIR" OR "interactive retrieval" OR "interactive IR" OR "user interaction" OR "interactive search" OR "information retrieval" OR "web search"*

Concepts	**Eligibility criteria for search and screening** **(* criteria used in Kelly & Sugimoto, 2013)**	**SPIDER** ***[PICO]***	**Draft query** (v8, 6/17/2022)
Assigned Tasks (C3)	Inclusion • The study should be experimental or quasi-experimental studies or studies in which researchers impose a simulated work task scenario and/or tasks. • Study setting could be in a lab or in the field/participant's normal setting. • Studies in which the researcher(s) assign tasks to study participants are included. Assigned tasks may involve search tasks or work tasks, as long as there is a search component. • Studies in which participants bring their own tasks to complete are included if the study also involves researcher-assigned tasks. • Studies in which researchers give queries to participants to use with the search system are included. Exclusion • Studies in which participants are asked to complete only self-generated tasks are excluded, unless they also include a simulated work task scenario and/or task. *	Phenomenon of Interest *[Intervention]*	*"scenario" OR "scenarios" OR "cover story" OR "cover stories" OR "simulated task" OR "simulated tasks" OR "synthetic task" OR "synthetic tasks" OR "imposed task" OR "imposed tasks" OR "imposed query" OR "imposed queries" OR lab OR "laboratory" OR "work task" OR "work tasks" OR "search task" OR "search tasks" OR "authentic task" OR "authentic tasks" OR "naturalistic task" OR "naturalistic tasks" OR "genuine task" OR "genuine tasks" OR "genuine information need" OR "genuine information needs" OR "authentic information need" OR "authentic information needs" OR "simulated work task" OR "simulated work tasks" OR "simulated information need" OR "simulated information needs"*
Empirical study (C4)	Inclusion • The study should be empirical and attempt to use at least some aspects of the scientific method. * • Experimental and quasi-experimental studies are explicitly included. • Observational studies that meet other criteria are also included.	Study Design *[Comparison]*	*"study" OR "studies" OR "experiment" OR "experiments" OR "quasi-experiment" OR "quasi-experiments" OR "evaluation" OR "evaluations"*

Concepts	**Eligibility criteria for search and screening** **(* criteria used in Kelly & Sugimoto, 2013)**	**SPIDER** ***[PICO]***	**Draft query** (v8, 6/17/2022)
Realism (C5)	Inclusion • For "realism search", only studies that mention realism or whether tasks are realistic will be included.	Evaluation *[Outcomes]*	*"realistic" OR "realism"*
	Inclusion • Quantitative, qualitative, and mixed methods are included.	Research Type	
	Inclusion • peer-reviewed publication, • written in English, • title + abstract available electronically, • published between 1997 and 2022. In 1997, the first of Borlund's papers discussing realism and simulated work tasks was published.		

6. Screening

The ST paper and realism paper (see sections below) screening and data extraction processes will be conducted in parallel using Covidence. Prior to that, a test Covidence instance with a subset of articles will be used to train and reduce the potential for bias of screeners. In the test Covidence, papers will be screened and resolved by all three researchers. The inclusion and exclusion criteria will be refined if necessary.

ST papers

For the set of "ST papers" (i.e., those that involve C1-4 whether or not they include C5), the retrieved citations from database searching (C1-4), citation chasing, and hand searching will be imported into EndNote and deduplicated.

Deduplicated citations will be uploaded to Covidence for ST review.

Two researchers will independently conduct title and abstract screening indicating whether the paper should be included in the review with three categories: clearly included ("yes"), clearly excluded ("no"), or unclear whether to include or exclude ("maybe"). The researchers will also tag the 'maybe' papers as such (e.g., 'unclear') so that those subset can be furthered verified as 'yes' or 'no'. Disagreements will be resolved by the third researcher (or consensus if only two researchers are working on the project).

Although not officially full text screening, the full text will be consulted to determine whether to include papers that have an "unclear" status after title and abstract screening OR, if needed, for papers with intractable conflicts.

As preliminary query testing suggests that we may have over 1500-2000 ST papers after deduplication, full text screening will only be conducted for the papers listed above.

Realism papers

The set of "realism papers" will consist of the deduplicated set of citations retrieved from database searching (C1-5) and those papers identified from the ST papers as mentioning realism. To identify ST papers that mention realism, we will conduct a full text search of the ST papers within EndNote for concept 5 query terms ("realism" or "realistic").

Deduplicated citations will be uploaded to the realism review in Covidence.

Two researchers will independently conduct title and abstract screening indicating whether the paper should be included in the review with three categories: clearly included ("yes"), clearly excluded ("no"), or unclear whether to include or exclude ("maybe"). Disagreements will be resolved by the third researcher (or consensus if only two researchers are working on the project).

Full text screening of realism papers will be independently completed by two researchers. Disagreement will also be resolved by the third researcher (or consensus if only two researchers are working on the project.

7. Data Extraction

Multiple rounds of data extraction are planned. As with screening, to reduce the potential for bias, the researchers who will extract data will be trained on the data extraction categories where definitions of the categories will be clarified if needed. To ensure that the data extractors have a shared understanding of the categories, all three researchers will extract data from a subset of the included articles, any disagreements will be discussed by all three researchers, and the data extraction categories, process, and tool (Excel spreadsheet) refined if necessary.

Our primary unit of analysis is the article or paper given the difficulty of connecting multiple studies (e.g., not all authors indicate whether a single study/data collection event is reported in multiple papers). As some papers report the results of multiple studies, it is likely that we will have nested data. In cases where the multiple studies use different sets of assigned tasks and/or discuss realism of each set of tasks, we will create multiple data extraction records (i.e., from each IIR study separately). If the multiple studies in a single paper use the same set of assigned tasks and do not discuss the realism of each study separately, we create one record for the paper. Our reporting will clearly indicate which level of data we are discussing (e.g., 150 studies reported in 120 papers).

Realism papers

To answer RQ1-4, data will be extracted from the "realism papers." The current version of the data extraction template is available as an appendix to this protocol. The final template is likely to undergo minor changes to optimize efficiency of data extraction (e.g., changing order, formatting).

Two researchers will independently extract data from realism papers that make it through the screening process. During data extraction, the researchers will extract data from the full text of each paper into an Excel spreadsheet.

Data extraction will take place in two parts: an initial round with a subset of articles and a final round. To prevent biased data extraction, a 25% random sample of records will be coded by both researchers, an agreement measure calculated, and any disagreements discussed and resolved by consensus. Then data extraction for the last 75% of the papers will take place using the same process.

We summarize the data to be extracted below separated by research question below. See the appendix for all data elements.

For all papers, some elements will be extracted that do not map to RQs.

- Identifier
- Number of eligible studies reported in paper

RQ1: For papers that use SWT and talk about realism: a) describe study including the (i) study participants, (ii) assigned tasks, (iii) study design, b) describe how they talk about realism (i) in general, (ii) of assigned tasks, and (iii) of participant-generated tasks.

a) Overview of studies

i) About study participants
   - Demographics: age, country, education, gender, affiliation, race
   - Sample size

ii) Assigned tasks
   - How included in paper: described, example in paper, full text in paper, full text in appendix
   - Description of assigned tasks: elements (indicative need, scenario, topic), level (work, search, information-seeking), type (exploratory, factual, decision-making, problem-solving, etc.)
   - Source of tasks: created SWT, re-used SWT, participant-generated task

iii) Study design
   - Study purpose: evaluate IR system(s) or features, understand search or work task behavior, evaluate measures, explore methodological issue
   - Sampling method: convenience, crowdsourcing, other
   - Data collection methods: questionnaires, interviews, system logs, think-aloud, eye-tracking
   - Search system: commercial search system, custom minimal search system, custom search system with novel elements, conversational search system, other
   - Device(s) used: laptop, desktop, mobile, tablet, conversational system
   - Study setting: lab, field, naturalistic
   - Study schedule: preset, participant-determined
   - Task- or topic-related measures: topic perceptions (interest, topic knowledge, difficulty), number of tasks completed
   - Other: time allowed to complete task, number of possible tasks, number of possible topics, training provided, pilot testing of tasks

b) How realism is described
   i) Type(s) of realism described: mundane, experimental, psychological, other
   ii) Realism: study design in general, SWT, participant-generated tasks
   iii) How realism is mentioned (open-coding)

RQ2: For papers that use SWT and talk about realism, how did they talk about steps they took to ensure realism of the assigned tasks (including tailoring) or the overall study design?

- Did they take steps to ensure realism?
- Steps taken to ensure realism: tailoring, pilot testing, __
- How SWTs tailored: to population, to setting, to system, other

RQ3: For papers that use SWT and talk about realism, how did they talk about how they assessed the realism of the assigned tasks or the overall study design.

- Did the study assess realism?
- What was the method used to assess realism?
- When was realism assessed during the study?

RQ4: For papers that use SWT and talk about realism, how did they talk about an impact or potential impact of realism on their results.

- Did study describe impact of realism?
- How did they describe potential impact?

ST papers

Data extraction for papers included in ST papers (but not in the set of realism papers) will focus on RQ1a and use the same set of elements as RQ1a as listed above.

8. Study Quality Assessment

We have designed our study to minimize biases and human error at multiple levels.

Publication bias

We will conduct a comprehensive literature search using five databases we have identified as the optimal combination with sufficiently different focuses and coverages. Database searches will be supplemented by hand-searching to remedy the partial coverage of publications in the five databases. We are also chasing citations to ensure the complete coverage of source titles under review. We believe our selection of the databases will guarantee the inclusion of all works that should be included in our reviews and help us avoid publication bias. While some guidelines for systematic reviews strongly encourage the inclusion of all types of grey literature, for the topic, we consider capturing all works appearing in relevant conference proceedings is sufficiently comprehensive for our systematic review.

Reviewer bias:

We aim to reduce reviewer bias in our data extraction. Therefore, we have developed and pilot-tested a standardized data extraction sheet with detailed instructions. Two or three review authors will extract data independently using the standardized data extraction sheet (included in the Appendix). Disagreements between the two authors will be resolved through discussion. The project lead will arbitrate if disagreements cannot be resolved by discussion. During the initial data extraction pilot phase, coding categories will be iteratively refined.

Risk of Bias Assessment using MMAT

As for the risk of bias assessment, we deemed the Mixed Method Appraisal Tool (MMAT) the best fit for our research questions. We plan to perform bias assessment on the subset of studies that use SWTs (C1-4) and discuss realism (C5). a subset of studies meeting the following conditions:

a. A study has to be included in RQ3 and RQ4 (i.e., assess realism).
b. For those selected in a, we will screen for those meeting the criteria, C1-4.
c. From b, we will select studies discussing realism or the impact of realism, which is our 5th criterion, C5.

Unlike the previous version, the latest version (2018) of MMAT advises against using the summative numeric score (i.e., single number). However, the MMAT creators allow users to use numeric scores on

the scale of 1-5 (5 being 100% quality criteria met and 1 being only 20% quality criteria met) as in the previous version if users strongly feel the need. However, we also feel that a summative numeric score conveys little information; thus, we have decided against using the numeric score as the quality measure.

9. Data Synthesis

Describe how you will analyze and summarize the included study results.

A combination of quantitative and qualitative data analysis will be conducted. Descriptive quantitative descriptive analysis will primarily be used. We will report the number of papers that are included in each of the research questions and the frequency of the closed codes for our data extraction.

A combination of qualitative thematic analysis and content analysis will be used to analyze the open-ended data collection.

10. Project Tools

Project tools include

- Citation managers: *SciWheel* and *EndNote*
- Citation chasing using *citationchaser* (Haddaway, Grainger, & Gray, 2021)
- Screening: Covidence
- Data extraction: Excel, Acrobat

11. Project Timetable

March – June 2022.

Preparation.

July 2022.

Finalize protocol.

Conduct searches

Pilot test eligibility criteria

Pilot test data collection

August – October 2022

Title and abstract screening (ST + realism papers)

Full text screening (realism papers)

Data extraction (realism papers)

Data analysis and results synthesis (realism papers)

Quality assessment (realism papers)

October 2022

Write manuscript with preliminary results

12. Research Team Member Roles

Describe the different tasks on the review and who will be responsible for what.

- Study protocol. All.
  - Draft, review, revise protocol. Anita, Lan, Yu Lee (study quality assessment, data extraction)
  - Contribute, review. Rebecca.
- Searches. All
  - Test queries. Anita
  - Conduct database searches. Rebecca
  - Citation chasing. Lan
  - Hand searching. TBD (was Yu Lee)
- Screening. Anita & Lan
- Data extraction. Anita & Lan
- Data analysis and results synthesis. Anita & Lan
- Quality assessment. Anita & Lan
- Write manuscript. All.
  - Methods section. Rebecca key contributor.
  - Full manuscript. Anita & Lan.

13. Protocol revision history

7/27/2022. First deposit in the Carolina Digital Repository. Any future changes will be detailed in this section.

14. References

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey and E. Aronson (Eds.), *The handbook of social psychology* (Vol. 2, pp. 1 – 79). Reading, MA: Addison - Wesley.

Aronson, E., Wilson, T. D., & Akert, R. M. (1994). *Social psychology: The heart and the mind*. New York: HarperCollins.

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 8-3. http://informationr.net/ir/8-3/paper152.html

Borlund, P. (2016). A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. *Journal of Documentation*. 72(3), 394-413. https://doi.org/10.1108/JD-06-2015-0068

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3), 225-50. https://doi.org/10.1108/EUM0000000007198

Capra, R., Velasco-Martin, J., & Sams, B. (2011). Collaborative information seeking by the numbers. In Proceedings of the 3rd International Workshop on Collaborative Information Retrieval - CIR ’11 (pp. 7–10). New York: ACM Press. https://doi.org/10.1145/2064075.2064078

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER Tool for qualitative evidence synthesis. Qualitative Health Research, 22(10), 1435–1443. https://doi.org/10.1177/1049732312452938

Crescenzi. A. (2019). Adaptation in information search and decision-making under time pressure. Ph.D. Dissertation. https://doi.org/10.17615/YT6K-AC37

Crescenzi, A. & Li, L. (2022). Assessing Realism in Simulated Work Tasks. In ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22). Association for Computing Machinery, New York, NY, USA, 266–271. https://doi.org/10.1145/3498366.3505831

Dabholkar, P. A. (1994). Incorporating choice into an attitudinal framework: Analyzing models of mental comparison processes. *Journal of Consumer Research*, 21(1), 100. https://doi.org/10.1086/209385

Darley, W. K., & Lim, J.-S. (1993). Assessing demand artifacts in consumer research: An alternative perspective. *Journal of Consumer Research.* 20(3), 489-495.

Haddaway, N. R., Grainger, M. J., Gray, C. T. (2021). citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching. https://doi.org/10.5281/zenodo.4543513

Hong, Q.N., Pluye, P, Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada. Retrieved from http://mixedmethodsappraisaltoolpublic.pbworks.com/w/page/24607821/FrontPage

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval, 3(1-2), 1–224. https://doi.org/10.1561/1500000012

Kelly, D. & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967-2006. Journal of the American Society for Information Science and Technology, 64(4), 745–770. https://doi.org/10.1002/asi.22799

Wildemuth, B.M., & Freund, L. (2009). Search tasks and their role in studies of search behaviors. Paper presented at HCIR 2009: Bridging Human-Computer Interaction and Information Retrieval, Washington, DC, October 23, 2009.

Wildemuth, B.M., & Freund, L. (2012). Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors. Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, New York, NY, 2012.

Wilson, T.D., Aronson, E., and Carlsmith, K. (2010). The art of laboratory experimentation. John Wiley & Sons, Inc., 51–81. https://doi.org/10.1002/9780470561119.socpsy001002