A Conceptual Framework for Implicit Evaluation of Conversational Search Interfaces

Abhishek Kaushik and Gareth J. F. Jones

ADAPT Centre, School of Computing Dublin City University, Ireland
abhishek.kaushik2@mail.dcu.ie, Gareth.Jones@dcu.ie

Abstract. Conversational search (CS) has recently become a significant focus of the information retrieval (IR) research community. Multiple studies have been conducted which explore the concept of conversational search. Understanding and advancing research in CS requires careful and detailed evaluation. Existing CS studies have been limited to evaluation based on simple user feedback on task completion. We propose a CS evaluation framework which includes multiple dimensions: search experience, knowledge gain, software usability, cognitive load and user experience, based on studies of conversational systems and IR. We introduce these evaluation criteria and propose their use in a framework for the evaluation of CS systems.

Keywords: Conversational Search · Evaluation · Human-Computer Search Interfaces

1 Introduction

Recent progress in artificial intelligence has brought tremendous advances in conversational systems and information retrieval (IR). This has led to increasing interest in Conversational Search (CS) using conversational engagement to complete IR tasks [29]. CS presents opportunities to support users in their search activities to improve the effectiveness and efficiency of information seeking, while reducing their cognitive load. A number of studies have been conducted to examine the concept of the CS [20, 29]. We believe that greater insight into the processes and potential of CS can be achieved using a detailed evaluation can help in advancing and understanding the exploitation of the paradigm of CS. These insights will help in enhancing proposed models and theories of CS.

This paper overviews the current methods and techniques used in the evaluation of conversational systems in different dimensions, and use them to define a framework for the definition and utilization of evaluation metrics for CS.

2 Background

This section introduces existing work examining the evaluation of interactive IR, conversational systems and CS.

2.1 Evaluation of Interactive Information Retrieval

Interactive IR (IIR) studies user interaction with search systems. The evaluation methods for IIR can be broadly classified into four major classes: contextual, interaction, performance and usability [21].

- **Contextual**: This measures the context in which a search and interaction activity occurs, and characterizes the subject and their information need. Characteristics of subjects include: age, sex, search experience, etc. Characteristics of information need focus on information seeking situations such as the subject's background knowledge, subject familiarity with the search topic, etc. These measures basically describe the context in which the information search occurs [8, 14, 21].
- **Interaction**: This focuses on characterizing the interactions between a search system and the user. This also includes the interactive search behaviour of the user, such as the length of each query, the number of queries entered, and the number of returned documents read, etc [21].
- **Performance**: This focuses on results obtained from the user's interaction with a search system, such as calculating precision, mean average precision, and recall of retrieved documents. According to Saracevic [31], these performance measures depend on the concept of relevance, the user's criteria of relevance assessment, and the techniques used for measuring the relevance.
- **User-feedback**: This captures the user's feelings and experiences of their interactions with the search system. This measure is also referred to as "usability" and is divided into multiple dimensions [15]. According to the International Organization for Standards (ISO), the key dimensions of usability are effectiveness, efficiency, and satisfaction [15].

2.2 Evaluation of Conversational Systems

Conversational engagement is currently being investigated as the mode of engagement for many human-machine applications. Research in conversational systems has proposed 6 dimensions for the evaluation of conversational agents [30, 42].

- **Extensive Capabilities**: Conversational systems should provide an error-free environment to the user [28, 30, 42]. This should include spell checking and auto-correction to support error free expression by the user of statements and questions. In addition, conversational systems should use an appropriate combination of multimedia and text content [10, 44].
- **User Interaction and Engagement**:This measure contains the following parameters: capability of initiating conversations, maintaining conversational engagement, identifying and distinguish target users, etc. [41].
- **Response speed**: The response speed of conversational agents should be sufficiently fast to prevent user frustration [30, 36, 42].

- **Functionality**: The overall capabilities of the system as measured qualitatively by users based on multiple variables such as the richness of media supported, navigation tools provided to support users, multi-modality of engagement, etc [28, 35].
- **Interoperability**: This defines the ability of a conversational system to exchange and make use of information. A standard conversational system enables the user to engage with multiple media channels [42].
- **Scalability**: This is a quantitative measure of the scope of the system to support multiple users. For example, the number of users supported by the conversational system at the same time, types of server that can accommodate the conversational system, database size, etc [42].

A well known framework for evaluation of conversational systems, is the Paradise framework [43]. This includes task success, conversation efficiency (task duration, dialogue turns), conversation quality (response accuracy), and user satisfaction (ease of the task, user behaviour). However, Paradise is of limited value for the evaluation of of conversational search systems, since it does not include important factors such as cognitive load and the knowledge gain during the search process. Moreover, Paradise focuses on a goal-oriented agent, which is different from a search-oriented task for which the requirements can change as the user progresses through the search process. Hence, Paradise is not suitable for evaluation of conversational search systems.

2.3 Evaluation of Conversational Search

In recent years, a number of studies have been conducted on CS systems and interfaces. There studies can be broken down into four different approaches: using existing conversational agents [22, 24], using human experts [37, 38], perceived experiment (Wizard-of-Oz approach), [3, 6, 27], and using rule-based or machine learning conversational interfaces [19]. Evaluation in most of this work is limited to NASA TASK Load [12] or SUS (Usability) [5] or both [4, 9]. Some of these studies have also investigated sentiment [9] in the user response to examine the relationship between the user's mood and task success.

There is also active research exploring the use of evaluation benchmarks for CS. Most notable is the TREC CAsT track [2, 7]. An alternative interpretation of CS is examined in the FIRE RCD track [18]. These tasks have examined query interpretation and response in the context of conversational engagement and query extraction from conversations respectively. In both cases, evaluation is largely limited to traditional approaches used for IR tasks.

Evaluation of interfaces in CS is a highly complex topic involving multiple dimensions including the user's background knowledge of the search topic, their familiarity with the conversational agent, etc. In this paper, we propose a framework for the evaluation of CS interfaces using five dimensions: user search experience [20], knowledge gain [45], cognitive and physical load [12], usability of the interface software [26] and user experience [13].

	Topics (0 (very low)- 7 (very high))
Search Formulation (Per-Search)	Background Knowledge
	Interest in Topic
	Anticipated Difficulty
Content Selection	Actual Difficulty
	Text Presentation Quality
	Average number of docs viewed per search
	The usefulness of Search results
	Text Relevance
Interaction with Content	Cognitively Engaged
	Suggestions Skills
	System Understanding Input
	Average Level of Satisfaction
Post Search	Search Success
	Presentation of the Search Results
	Expansion of knowledge after the search
	Understanding about the Topic

Table 1. Flowchart of characteristics of the search process [39] by change in knowledge structure

3 Framework for the Implicit Evaluation of Conversational Search Interfaces

In this section we introduce our framework for the evaluation of conversational search with focuses on the multiple dimensions relating to the user. An approach of this sort is also advocated in the summary report from the Dagstuhl Seminar on conversational search [1].

Most CS studies so far reported have focused on user search experience of the task or the usability of CS systems. This has provided feedback focused on user search experience. In our framework, evaluation is based on five factors responsible for the needs of CS outlined in the next section.

3.1 Essential Factors for Conversational Search

We identify the following essential factors for the evaluation of CS interfaces.

1. Cognitive Load: Conventional search can impose a significant cognitive load on the searcher [17]. An important factor in the evaluation of conversational systems is measurement of the cognitive load experienced by users while using the system.
2. Cognitive Engagement: It has been observed that users get frustrated if they find it difficult to search about their topic of interest to satisfy their information need. Frustration can reduce the user's engagement with a search system and their associated effort to locate relevant information.
3. Search as Learning: Learning while searching is an integral part of the information seeking process. In our current study, we propose a metric which

breaks this down into three phases [20,39], as shown in Table 1. It has been observed in conventional search that high cognitive load and lower cognitive engagement impacts on user learning during the search process [11].

4. Knowledge Gain: Satisfaction of the user's information need is directly related to their knowledge gain about the search topic. Knowledge gain can be measured based on recall of new facts gained after the completion of the search process [45].
5. User Experience (UX): Another important aspect that needs to be considered for evaluation of CS systems is UX. User experience is generally classified into two aspects: pragmatic and hedonic, which can be further divided into six components: attractiveness, perspicuity, efficiency, dependability, stimulation and novelty [13]. These factors all provide measures of user ease of use and the dependability of a conversational system [13].
6. Software Usability: CS studies generally do not explore the dimensions of software usability. However, it is important to understand the challenges and opportunities of conversational systems on the basis of software requirements analysis. This allows a system to be evaluated based on real life deployment and to identify areas for improvement. Lower effectiveness and efficiency of a software system can increase cognitive load, reduce engagement and act as a barrier in the process of learning while searching.

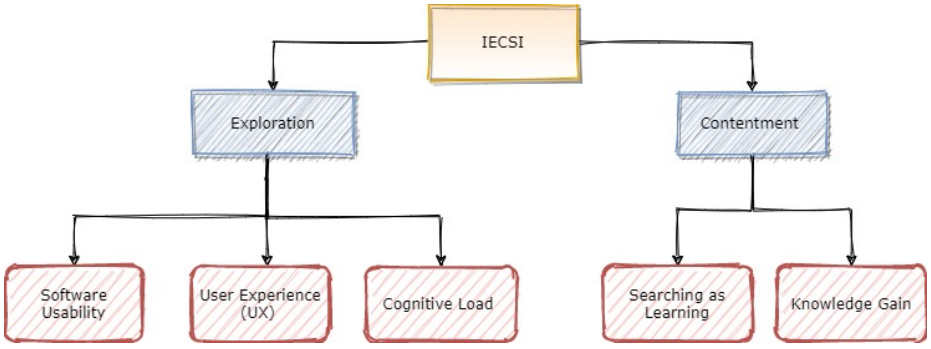

Fig. 1. Implicit Evaluation of a Conversational Search Interface (IECSI)

3.2 Designing our Conceptual Framework

In our work on CS, we have conducted multiple studies of CS and the use of associated conversational agents. In conjunction with this, we have also conducted four user studies on conventional IR systems, a commercial conversational system (Alexa Echo Show) and conversational search interfaces for complex search tasks. In these studies, we have examined user behaviour and user expectations with respect to CS. Based on our investigations [2, 17–20], we propose an

evaluation framework for CS interfaces. The framework is divided into two segments: Exploration and Contentment, as shown in Figure 1. This section outlines the combination of standard questionnaires of multiple dimensions to form our proposed Implicit Evaluation for Conversational Search Interface (IECSI). The details are as follows:

Explore Segment: This segment focuses on exploring and experiencing CS interfaces, and is classified into three components: Software Usability, User Experience, and Cognitive Load.

1. Software Usability: Usability is an important consideration for the evaluation of interactive software. the IBM Computer Usability Satisfaction Questionnaires enables psychometric evaluation from the perspective of the user, and is known as the Post-Study System Usability Questionnaire (PSSUQ) [26]. The PSSUQ includes four dimensions: overall satisfaction score (OVERALL), system usefulness (SYSUSE), information quality (INFOQUAL) and interface quality (INTERQUAL), which includes sixteen parameters.
2. User Experience: UX is measured using a questionnaire for interactive product known as the User Experience Questionnaire (UEQ-S) [13, 25, 34]. This questionnaire also enables us to analyse and interpret outcomes by comparing against a benchmark dataset of outcomes for other interactive products. This questionnaire also provides us with the opportunity to compare interactive products with each other. UEQ-S contains two meta quality dimensions: pragmatic and hedonic. Each dimension contains four different parameters as shown in the Table 2. Pragmatic quality explores the usage experience of a conversational search system. Hedonic quality explores the pleasantness of using the system.
3. Cognitive Load: An important consideration in the evaluation of CS interfaces is their impact on the user's cogntive load during the search process. To measure the user's workload, the NASA Ames Research Centre proposed the NASA Task Load Index [12]. This is a multi-dimensional rating procedure which provides a measurement of the overall workload during the process or event. This workload is classified into six subscales: mental, physical, temporal, own performance, effort and frustration. Out of these six dimensions, three are related to the demand imposed on the subject due to the task (mental, physical and temporal) and the remaining three to the interaction of the subject with the system (effort, frustration and performance). This implicit evaluation enables us to examine the cognitive load and cognitive engagement of the user while using a system.

Contentment Segment: This segment focuses on information need satisfaction during the search process. It includes a questionnaire based on interaction while searching, learning during searching and knowledge gain arising from the search activity:

	Negative	1 2 3 4 5 6 7	Positive
Pragmatic quality	obstructive	□□□□□□□	supportive
	complicated	□□□□□□□	easy
	inefficient	□□□□□□□	efficient
	confusing	□□□□□□□	clear
Hedonic quality	boring	□□□□□□□	exciting
	not interesting	□□□□□□□	interesting
	conventional	□□□□□□□	inventive
	usual	□□□□□□□	leading edge

Table 2. Scales pragmatic quality and hedonic quality

Parameter	Definition
Dqual	Comparison of the quality of facts in the summary in range 0-3 where 0 represents irrelevant facts and 3 specific details with relevant facts.
Dintrp	Measures the association of facts in a summary in the range 0-2 where 0 represents no association of the facts and 2 that all facts in a summary are associated with each other in a meaning.
Dcrit	Examines the quality of critiques of topic written by the author in range the 0-1 where 0 represents facts are listed without analysis and 1 where both advantages and disadvantages of the facts are given.

Table 3. Summary Comparison Metric [45]

1. Search as Learning: As discussed in Section 3.1, it is important to observe whether a CS system supports the user effectively in their engagement with the search system, and enables the user's knowledge gain arising from the search process. To better understand this process, we decided to separately measure the factors of user interaction and modification of their mental knowledge structures. We developed a questionnaire [20], as shown in Table 1, to observe user interaction behaviour, inspired by Vakkeri's model of search as learning [39, 40].
2. Knowledge Gain: To measure the knowledge gain, the user is required to write a pre-search summary and a post-search summary relating to the search topic. This summary is manually evaluated by an independent assessor on three sub dimensions: Quality of Facts (Dqual), Interpretation (Dintrp) and Critiques (Dcrit), as shown in Table 3 [45].

3.3 Developing the Evaluation Process

The overall evaluation process is shown in the Figure 2. The user completes a pre-search questionnaire and then a post-search questionnaire. The questionnaire is based on the metric discussed in Section 3.2, and as shown in Figure 3. To maintain uniformity, subjects rate each parameter on a 7-point Likert scale [16], where the scale ranges from 1 (very low) to 7 (very high) on each questionnaire. The evaluation is conducted from two perspectives: a) comparison of a

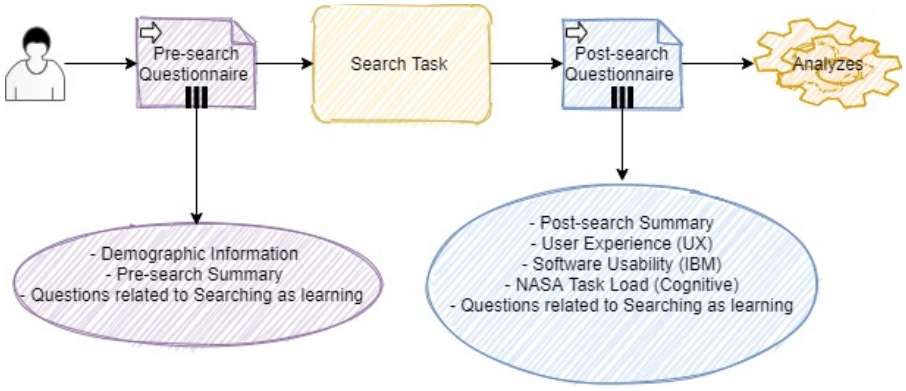

Fig. 2. Evaluation Process

conversational interface with a conventional search system, b) evaluating only a conversational interface based on a provided benchmark. Details are given below:

- Comparison of conversational interface with conventional search system: This evaluation method enables comparison of conventional and conversational search interfaces based on 5 dimension metrics, as discussed in Section 3.2. The user completes two search tasks one each using each search setting (conventional and conversational). For each task, the user completes a pre-search questionnaire and a post search questionnaire. This analysis is intended to provide better insights into the operation of a CS system and contrasting user opinions of each type of interface.
- Evaluating the conversational interface based on standard numerical benchmarks: Most of the metrics introduced in the framework have a standard numerical benchmark [**?**, 12, 20, 26, 45]. Pre-search and post-search questionnaire scores for each dimension of evaluation can be compared using their standard benchmarks. This not only provides an estimation of evaluation, but also provides an opportunity to explore the conversational interface with the standard system benchmark. Furthermore, this allows us to understand user expectation in general in all dimensions. This provides empirical measurability of a conversational search interface in the light of the benchmark. Moreover, this can help us to understand how far or close the current CS interface is to the user's expectations.

As per above the prospective, it is very important to analyse the data critically including use of statistical significance tests. If the results are significant, this can be used to develop a separate benchmark for the CS interface to assist other researchers in comparing their studies on CS search interfaces.

3.4 Implementation and Analysis of the Framework

As mentioned earlier, the user is required to complete pre-search and post search questionnaires. We have developed these questionnaires by combining the di-

mensions mentioned earlier in Section 3.2[1]. The details of the pre-search and post-search questionnaires are described below.

1. **Pre-search Questionnaire:** This focuses only on contentment, and contains questions on demographic details of searcher, background knowledge of the searcher about the search topic, interest in the search topic, searcher experience of using conversational system, etc.
2. **Post-search Questionnaire:** This focuses on contentment and exploration, and contains questions on knowledge gain after search, based on interactions (e.g., How many documents reviewed by user?), software usability, UX, cognitive load, etc.

	Metric	Pre-Search	Post-Search	Evaluation
Contentment	Searching as learning	User will be asked to fill the Questionnaire after reading the search task. This is to understand the state of user in the phase of search formulations in search process as shown in the Table 2.	User will be asked to fill the Questionnaire post search task. This is to understand the state of user in the phase of Content Selection, Interaction as shown in the Table 2. This questionnaire will also allow user to provide direct and indirect feedback about the search success, expansion of knowledge and interaction experience	• To keep the uniformity, subjects can rate each parameter on a 7-point Likert scale where 1 (very low) and 7 (very high) on each questionnaire. • Some parameters are subjective
	Knowledge Expansion	User will write a summary based on Pre-existing knowledge on search topic	User will write a summary on search topic based on knowledge gain after search	Summary will be coded by at least two independent assessors (Kappa Coefficient >.85) based on three categories such as Quality, Interpretation and Critiques as shown in Table 2.
Exploration	Software Usability		User will fill the Psychometric Evaluation for software from the perspective of the user By IBM discussed in tool [19]	• Evaluated using Sixteen as mentioned in Table (Likert scale 0-7). • To verify your system is as per the standard, the overall average is greater than 3 [25]
	User Experience (UX)		User will fill UEQ-S to provide feedback to measure the user experience on interactive system as shown in Figure 2.	• Evaluated based on UEQ standard metric tool. • Can also compared with the standard benchmark by using tool [26]
	Cognitive Load		User will ask to fill the NASA Task Load Index (TLX) questions assesses workload [8].	• Likert scale (1 to 7) • Can be evaluated based on the mean value. • Can be compared with the other system

Fig. 3. Implicit Evaluation for Conversational Search Interface Metric

Each question is evaluated based on a Likart score (0,7), except for the knowledge gain metric. As described, the framework is classified into two sections: exploration and contentment, as shown in Figure 3. Details are as follows:

1. Exploration: The questionnaire to investigate exploration is aligned to the user based on their search experience. As such, the conversational interface is evaluated based on the post-search questionnaire. The mean score of each question is calculated based on the number of users. Analysis is conducted using both Quantitative Analysis and Qualitative Analysis.

[1] The questionnaires can be found at https://forms.gle/MaooazzEfQJ4sTpPA

(a) Quantitative Analysis: This is based on the mean score of the participants in the study, statistical testing is carried out based on the population and nature of the experiment. When comparing a conventional system and a conversational system, we are able to perform dependent significant testing, since the population undertaking the experiment in both settings is the same. And, if we are comparing the mean score of the conversational interface with a standard benchmark, we can conduct independent significance testing. This statistical testing enables us to understand how systems differ. Additionally, each dimension discussed above in Section 3.2 has a standard tool for analysis.

(b) Qualitative Analysis: The different dimensions are annotated based on comparison of the mean values for the study participants. A mean value between 2 and 4 represents a neutral evaluation of the corresponding scale (yellow dimension), a mean > 4 represents a positive evaluation (green dimension) and mean < 2 represents a negative evaluation (blue dimension). After comparing the mean, each question is annotated based on the dimensions. The dimensions are annotated by two independent analysts with the Kappa coefficient (Approx .85), then the dimension is counted for each section such as software usability, user experience [32, 33], cognitive load. As per the dimension, the section of the interface that need to be improved can be identified. For example, if software usability gets more red dimensions, then the interface needs to be improved with respect to software usability.

2. Contentment: The questionnaire to investigate contentment is aligned to the user's pre-search knowledge and post-search knowledge. As discussed earlier, contentment evaluation is designed to investigate user learning while searching, and their knowledge expansion arising from the search process. The analysis can again be conducted using both Quantitative Analysis and Qualitative Analysis.

(a) Quantitative Analysis: Based on the mean score for the study participants of searching as learning and knowledge gain (the difference between post search and pre-search summaries of each setting (conventional system and conversational system)) parameters, statistical testing can be applied.

(b) Qualitative Analysis: Search as learning questions, as shown in Table 1, are annotated, evaluated and analyzed based on different dimensions as discussed in the exploration Qualitative Analysis section. Pre-search and Post-search summaries can be compared based on the parameters discussed in Table 3. The summary is scored against all these factors by two independent analysts with the Kappa coefficient (Approx 0.85) [23]. For each parameter, the difference between pre-search and post search summaries is calculated. If the difference of Dqual> 1.5, Dintrp > 1 and Dcrit > 0. it is assumed that the user has increased their knowledge by more than 50%.

4 Concluding Remarks

The concept of conversational search (CS) remains an ongoing topic of research. A crucial part of this work is evaluation of CS. Studies of CS to date have mainly been based on user experience. This overlooks interaction with the system and changes in the user's knowledge structure. In this paper, we examine the factors that can impact on the effectiveness of a CS interface. Following our investigation, we propose an evaluation framework for CS that incorporates the evaluation methods from interactive IR and conversational systems. We believe that our proposed evaluation framework for CS to be practical and applicable in real life scenarios, and will provide greater insights for understanding and advancing CS processes than is possible using the evaluation methods used in existing work on CS. We are currently working on the validation of our proposed framework within our ongoing study of CS.

Acknowledgement

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

References

1. Anand, A., Cavedon, L., Joho, H., Sanderson, M., Stein, B.: Conversational Search (Dagstuhl Seminar 19461). Dagstuhl Reports **9**(11), 34–83 (2020). https://doi.org/10.4230/DagRep.9.11.34, https://drops.dagstuhl.de/opus/volltexte/2020/11983
2. Arora, P., Kaushik, A., Jones, G.J.F.: DCU at the TREC 2019 conversational assistance track. In: Proceedings of TREC 2019 (2020)
3. Avula, S.: Wizard of oz: Protocols and challenges in studying searchbots to support collaborative search
4. Avula, S., Arguello, J., Capra, R., Dodson, J., Huang, Y., Radlinski, F.: Embedding search into a conversational platform to support collaborative search. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. pp. 15–23 (2019)
5. Brooke, J.: Sus: a "quick and dirty'usability. Usability evaluation in industry p. 189 (1996)
6. Dahlbäck, N., Jönsson, A., Ahrenberg, L.: Wizard of Oz studies: why and how. In: Proceedings of the 1st international conference on Intelligent user interfaces. pp. 193–200 (1993)
7. Dalton, J., Xiong, C., Callan, J.: CAsT 2019: The Conversational Assistance Track overview. In: Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC. pp. 13–15 (2019)
8. Dourish, P.: What we talk about when we talk about context. Personal and ubiquitous computing **8**(1), 19–30 (2004)
9. Dubiel, M., Halvey, M., Azzopardi, L., Daronnat, S.: Investigating how conversational search agents affect user's behaviour, performance and search experience. In: The second international workshop on conversational approaches to information retrieval (2018)

10. Eeuwen, M.v.: Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers. Master's thesis, University of Twente (2017)
11. Gwizdka, J.: Distribution of cognitive load in web search. Journal of the American Society for Information Science and Technology **61**(11), 2167–2187 (2010)
12. Hart, S., Staveland, L.: Development of nasa-tlx (task load index): Results and theoretical research, human mental workload (1988)
13. Hinderks, A., Schrepp, M., Thomaschewski, J.: A benchmark for the short version of the user experience questionnaire. In: WEBIST. pp. 373–377 (2018)
14. Ingwersen, P.: Järvelin. k.(2005b). the turn: Integration of information seeking and retrieval in context (2005)
15. Iso, W.: 9241-11. ergonomic requirements for office work with visual display terminals (vdts). The international organization for standardization **45**(9) (1998)
16. Joshi, A., Kale, S., Chandel, S., Pal, D.K.: Likert scale: Explored and explained. Current Journal of Applied Science and Technology pp. 396–403 (2015)
17. Kaushik, A.: Dialogue-based information retrieval. In: European Conference on Information Retrieval. pp. 364–368. Springer (2019)
18. Kaushik, A., Bhat Ramachandra, V., Jones, G.J.F.: DCU at the FIRE 2020 Retrieval from Conversational Dialogues (RCD) task. In: FIRE 2020 proceeding (2019)
19. Kaushik, A., Bhat Ramachandra, V., Jones, G.J.F.: An interface for agent supported conversational search. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR 2020). pp. 452–456 (2020)
20. Kaushik, A., Jones, G.J.F.: Exploring current user web search behaviours in analysis tasks to be supported in conversational search. In: Second International Workshop on Conversational Approaches to Information Retrieval (CAIR'18), July 12, 2018, Ann Arbor Michigan, USA (2018)
21. Kelly, D., et al.: Methods for evaluating interactive information retrieval systems with users. Foundations and Trends® in Information Retrieval **3**(1–2), 1–224 (2009)
22. Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M.: Toward voice query clarification. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 1257–1260. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3210160, https://doi.org/10.1145/3209978.3210160
23. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
24. Landoni, M., Matteri, D., Murgia, E., Huibers, T., Pera, M.S.: Sonny, cerca! evaluating the impact of using a vocal assistant to search at school. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 101–113. Springer (2019)
25. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Symposium of the Austrian HCI and usability engineering group. pp. 63–76. Springer (2008)
26. Lewis, J.R.: Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction **7**(1), 57–78 (1995)
27. McTear, M., Callejas, Z., Griol, D.: Evaluating the conversational interface. In: The Conversational Interface, pp. 379–402. Springer (2016)

28. Morrissey, K., Kirakowski, J.: 'realness' in chatbots: Establishing quantifiable criteria. In: International Conference on Human-Computer Interaction. pp. 87–96. Springer (2013)
29. Radlinski, F., Craswell, N.: A theoretical framework for conversational search. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 117–126. ACM (2017)
30. Radziwill, N.M., Benton, M.C.: Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579 (2017)
31. Saracevic, T.: Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for information science **26**(6), 321–343 (1975)
32. Sauro, J., Lewis, J.R.: Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann (2016)
33. Schrepp, M.: User experience questionnaire (2018), https://www.ueqonline.org/ accessed on 10-01-2021
34. Schrepp, M., Hinderks, A., Thomaschewski, J.: Design and evaluation of a short version of the user experience questionnaire (ueq-s). IJIMAI **4**(6), 103–108 (2017)
35. Staven, T.: What makes a good bot or not? (2017), https://www.unit4.com/blog/2017/03/whatmakesagoodbotornot accessed on 20-02-2019
36. Thieltges, A., Schmidt, F., Hegelich, S.: The devil's triangle: Ethical considerations on developing bot detection methods. In: 2016 AAAI Spring Symposium Series (2016)
37. Trippas, J.R., Spina, D., Cavedon, L., Sanderson, M.: How do people interact in conversational speech-only search tasks: A preliminary analysis. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. pp. 325–328 (2017)
38. Trippas, J.R., Spina, D., Thomas, P., Sanderson, M., Joho, H., Cavedon, L.: Towards a model for spoken conversational search. Information Processing & Management **57**(2), 102162 (2020). https://doi.org/https://doi.org/10.1016/j.ipm.2019.102162
39. Vakkari, P.: Searching as learning: A systematization based on literature. Journal of Information Science **42**(1), 7–18 (2016)
40. Vakkari, P., Pennanen, M., Serola, S.: Changes of search terms and tactics while writing a research proposal: A longitudinal case study. Information processing & management **39**(3), 445–463 (2003)
41. Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., Goel, R., Yang, S., Raju, A.: On evaluating and comparing conversational agents. CoRR **abs/1801.03625** (2018), http://arxiv.org/abs/1801.03625
42. Vyas, B.: 6 key metrics to measure the performance of your chatbot (2017), https://chatbotslife.com/6-key-metrics-to-measure-the-performance-of-your-chatbot-5fd0adfd0b5b accessed on 20-02-2019
43. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: Paradise: A framework for evaluating spoken dialogue agents. arXiv preprint cmp-lg/9704004 (1997)
44. Wilson, H.J., Daugherty, P., Bianzino, N.: The jobs that artificial intelligence will create. MIT Sloan Management Review **58**(4), 14 (2017)
45. Wilson, M.J., Wilson, M.L.: A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. Journal of the Association for Information Science and Technology **64**(2), 291–306 (2013)