Foundations and Trends® in
Information Retrieval
Vol. 3, Nos. 1–2 (2009) 1–224

DOI: 10.1561/1500000012

Methods for Evaluating Interactive Information Retrieval Systems with Users

By Diane Kelly

Contents

Foundations and Trends® in
Information Retrieval
Vol. 3, Nos. 1–2 (2009) 1–224

DOI: 10.1561/1500000012

Methods for Evaluating Interactive Information Retrieval Systems with Users

Diane Kelly

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, dianek@email.unc.edu

Abstract

This paper provides overview and instruction regarding the evaluation of interactive information retrieval systems with users. The primary goal of this article is to catalog and compile material related to this topic into a single source. This article (1) provides historical background on the development of user-centered approaches to the evaluation of interactive information retrieval systems; (2) describes the major components of interactive information retrieval system evaluation; (3) describes different experimental designs and sampling strategies; (4) presents core instruments and data collection techniques and measures; (5) explains basic data analysis techniques; and (4) reviews and discusses previous studies. This article also discusses validity and reliability issues with respect to both measures and methods, presents background information on research ethics and discusses some ethical issues which are specific to studies of interactive information retrieval (IIR). Finally, this article concludes with a discussion of outstanding challenges and future research directions.

1

Introduction

Information retrieval (IR) has experienced huge growth in the past decade as increasing numbers and types of information systems are being developed for end-users. The incorporation of users into IR system evaluation and the study of users' information search behaviors and interactions have been identified as important concerns for IR researchers [46]. While the study of IR systems has a prescribed and dominant evaluation method that can be traced back to the Cranfield studies [54], studies of users and their interactions with information systems do not have well-established methods. For those interested in evaluating interactive information retrieval systems with users, it can be difficult to determine how to proceed from a scan of the literature since guidelines for designing and conducting such studies are for the most part missing.

In interactive information retrieval (IIR), users are typically studied along with their interactions with systems and information. While classic IR studies abstract humans out of the evaluation model, IIR focuses on users' behaviors and experiences — including physical, cognitive and affective — and the interactions that occur between users and systems, and users and information. In simple terms, classic IR evaluation asks

the question, does this system retrieve relevant documents? IIR evaluation asks the question, can people use this system to retrieve relevant documents? IIR studies include both system evaluations as well as more focused studies of users' information search behaviors and their interactions with systems and information. IIR is informed by many fields including traditional IR, information and library science, psychology, and human–computer interaction (HCI). IIR has often been presented more generally as a combination of IR and HCI, or as a sub-area of HCI, but Ruthven [225] argues convincingly that IIR is a distinct research area. Recently, there has been interest in HCIR, or human computer information retrieval, but this looks similar to IIR and papers about this area have not established its uniqueness (e.g., [191]).

The proposition that IR systems are fundamentally interactive and should be evaluated from the perspective of users is not new. A review of IR literature reveals that many leaders in the field were writing about and studying interactive IR systems during the early years of IR research. For instance, Salton wrote a paper entitled "*Evaluation problems in interactive information retrieval*" which was published in 1970. In this paper, Salton [229] identified user effort measures as important components of IR evaluation, including the attitudes and perceptions of users. Cleverdon et al. [55] identified presentation issues and user effort as important evaluation measures for IR systems, along with recall and precision. Tague and Schultz [259] discuss the notion of user friendliness.

Some of the first types of IR interactions were associated with relevance feedback. Looking closely at this seemingly simple type of interaction, we see the difficulties inherent in IIR studies. Assuming that users are provided with information needs, each user is likely to enter a different query, which will lead to different search results and different opportunities for relevance feedback. Each user, in turn, will provide different amounts of feedback, which will create new lists of search results. Furthermore, causes and consequences of these interactions cannot be observed easily since much of this exists in the user's head. The actions that are available for observation — querying, saving a document, providing relevance feedback — are surrogates of cognitive activities. From such observable behaviors we must *infer* cognitive

activity; for instance, users who save a document may do so because it changes or adds to their understanding of their information needs.

User–system interactions are influenced by a number of other factors that are neither easily observable nor measurable. Each individual user has a different cognitive composition and behavioral disposition. Users vary according to all sorts of factors including how much they know about particular topics, how motivated they are to search, how much they know about searching, how much they know about the particular work or search task they need to complete, and even their expectations and perceptions of the IIR study [139, 194]. Individual variations in these factors mean that it is difficult to create an experimental situation that all people will experience the same, which in turn, makes it difficult to establish causal relationships. Moreover, measuring these factors is not always practical since there are likely a large number of factors and no established measurement practices.

The inclusion of users into any study necessarily makes IIR, in part, a behavioral science. As a result, appropriate methods for studying interactive IR systems must unite research traditions in two sciences which can be challenging. It is also the case that different systems, interfaces and use scenarios call for different methods and metrics, and studies of behavior and interaction suggest research designs that go beyond evaluation. For these reasons, there is no strong evaluation or experimental framework for IIR evaluations as there is for IR studies. IIR researchers are able to make many choices about how to design and conduct their evaluations, but there is little guidance about how to do this.

1.1 Purpose and Scope

There is a small body of research on evaluation models, methods, and metrics for IIR, but such studies are the exception rather than the rule (e.g., [34, 149]). In contrast to other disciplines where studies of methods and experimental design comprise an important portion of the literature, there are few, if any, research programs in IIR that investigate these issues and there is little formal guidance about how to conduct such studies, despite a long-standing call for such work

[231]. Tague's [260, 262] work and select chapters of the edited volume by Spärck-Jones [246] provide good starting points, but these writings are 15–20-years-old. While it might be argued that Spärck-Jones' book still describes the basic methodology behind traditional IR evaluations, Tague's work, which focuses on user-centered methods, needs updating given changes in search environments, tasks, users, and measures. It is also the case that Tague's work does not discuss data analysis. One might consult a statistics textbook for this type of information, but it can sometimes be difficult to develop a solid understanding of these topics unless they are discussed within the context of one's own area of study.

The purpose of this paper is to provide a foundation on which those new to IIR can make more informed choices about how to design and conduct IIR evaluations with human subjects.[1] The primary goal is to catalog and compile material related to the IIR evaluation method into a single source. This paper proposes some guidelines for conducting one basic type of IIR study — laboratory evaluations of experimental IIR systems. This is a particular kind of IIR study, but not the only kind. This paper is also focused more on quantitative methods, rather than qualitative. This is not a statement of value or importance, but a choice necessary to maintain a reasonable scope for this paper.

This article does not prescribe a step-by-step recipe for conducting IIR evaluations. The design of IIR studies is not a linear process and it would be imprudent to present the design process in this way. Typically, method design occurs iteratively, over time. Design decisions are interdependent; each choice impacts other choices. Understanding the possibilities and limitations of different design choices help one make better decisions, but there is no single method that is appropriate for all study situations. Part of the intellectual aspects of IIR is the method design itself. Prescriptive methods imply research can only be done in

[1] The terms *user* and *subject* are often used interchangeably in published IIR studies. A distinction between these terms will be made in Section 7. Since this paper focuses primarily on laboratory evaluations, the term *subject* will be used when discussing issues related to laboratory evaluations and *user* will be used when discussing general issues related to all IIR studies. *Subject* is used to indicate a person who has been sampled from the *user* population to be included in a study.

one way and often prevent researchers from discovering better ways of doing things.

The focus of this paper is on text retrieval systems. The basic methodological issues presented in this paper are relevant to other types of IIR systems, but each type of IIR system will likely introduce its own special considerations and issues. Additional attention is given to the study of different types of IIR systems in the final section of this paper. Digital libraries, a specific setting where IIR occurs, are also not discussed explicitly, but again, much of the material in this paper will be relevant to those working in this area [29].

Finally, this paper surveys some of the work that has been conducted in IIR. The survey is not intended to be comprehensive. Many of the studies that are cited are used to illustrate particular evaluation issues, rather than to reflect the state-of-the-art in IIR. For a current survey of research in IIR, see Ruthven [225]. For a more historic perspective, see Belkin and Vickery [23].

1.2 Sources and Recommended Readings

A number of papers about evaluation have been consulted in the creation of this paper and have otherwise greatly influenced the content of this paper. As mentioned earlier, the works of Tague [260, 262, 263, 264] and Tague and Schultz [259] are seminal pieces. The edited volume by Spärck-Jones [246] also formed a foundation for this paper.

Other research devoted to the study and development of individual components or models for IIR evaluation have also influenced this paper. Borlund [32, 34] has contributed much to IIR evaluation with her studies of simulated information needs and evaluation measures. Haas and Kraft [115] reviewed traditional experimental designs and related these to information science research. Ingwersen and Järvelin [139] present a general discussion of methods used in information seeking and retrieval research. Finally, the TREC Interactive Track [80] and all of the participants in this Track over the years have made significant contributions to the development of an IIR evaluation framework.

Review articles have been written about many topics discussed in this paper. These articles include Sugar's [255] review of user-centered

perspectives in IR and Turtle et al.'s [277] review of interactive IR research as well as Ruthven's [225] more recent version. The *Annual Review of Information Science and Technology* (*ARIST*) has also published many chapters on evaluation over its 40-year history including King's [173] article on the design and evaluation of information systems,[2] Kantor's [161] review of feedback and its evaluation in IR, Rorvig's [223] review of psychometric measurement in IR, Harter and Hert's [123] review of IR system evaluation, and Wang's [290] review of methodologies and methods for user behavior research.

Several special issues of journals about evaluation of IR and IIR systems are also worth mentioning. The most current is Borlund and Ruthven's [37] special issue of *IP&M* about evaluating IIR systems. Other special issues include Dunlop et al.'s [82] special issue of *Interacting with Computers* and Harman's [120] special issue of *IP&M*, which included Robertson and Hancock-Beaulieu's [221] discussion of changes in IR evaluation as a result of new understandings of relevance, interaction and information behavior. These articles, along with Savage-Knepshield and Belkin's [240] analysis of how IR interaction has changed over time, Saracevic's [233] assessment of evaluation in IR, and Ingwersen and Järvelin's [139] book on information seeking and retrieval are great background reading for those interested in the evolution of IIR systems and evaluation.

In addition to the sources from the IIR and IR literature, a number of sources related to experimental design and statistics were instrumental in the development of this paper: Babbie [13], Cohen [56], Gravetter and Wallnau [110], Myers and Well [200], Pedhazur and Schmelkin [208], and Williams [296].

1.3 Outline of Paper

The paper begins with a description of IIR and short discussion of its history. The next section reviews general approaches to studying IIR. Although this paper focuses on laboratory evaluations, other approaches are discussed briefly. Section 5 introduces

[2] Six articles were published in *ARIST* with the title, *Design and evaluation of information systems*, during the period 1968–1975.

research basics — research questions, theory, hypotheses, and variables. More advanced readers might want to skip this section, although the discussion of levels of measurement is particularly important for understanding the later material on statistics. Basic experimental designs are introduced in Section 6, followed by a discussion of sampling (Section 7). Instruments and data collection techniques are then presented in Section 8, followed by a discussion of some of the more common measures used in IIR evaluation (Section 10). A lengthy section on data analysis is in Section 11; although some instruction regarding qualitative data analysis is provided, this section primarily focuses on quantitative data analysis. This presentation starts with the basics of statistical data analysis, so advanced readers might want to skim parts of this section. Discussions of validity and reliability and research ethics are in Section 12. The paper concludes with future directions and challenges in Section 14.

2

What is IIR?

What is meant by *IIR*? An easy answer is that IIR is IR with users, but this does not really tell the whole story. One way to think about IIR is to place it in the middle of a continuum that is anchored by system focused studies and human focused studies (Figure 2.1). Studies situated at the system end of the spectrum are focused on developing and evaluating retrieval algorithms and indexing techniques. Studies such as those conducted in most TREC tracks would be examples of studies at this end of the continuum. There are no real users in these types of studies. Assessors may be used to create topics and evaluate documents, but they do not really function as users *per se*. Studies at the system end of the continuum can also be characterized by a lack of interaction — even if assessors are present, no searching takes place. Voorhees and Harman's [288] edited book describing TREC can be consulted for examples of these types of studies.

As we move along the continuum, the next type of study we observe are those that employ users to make relevance assessments of documents in relation to tasks. Users are basically used to build infrastructure so that a system-oriented study can be conducted. No searching is conducted and there is usually a lack of interest in users'

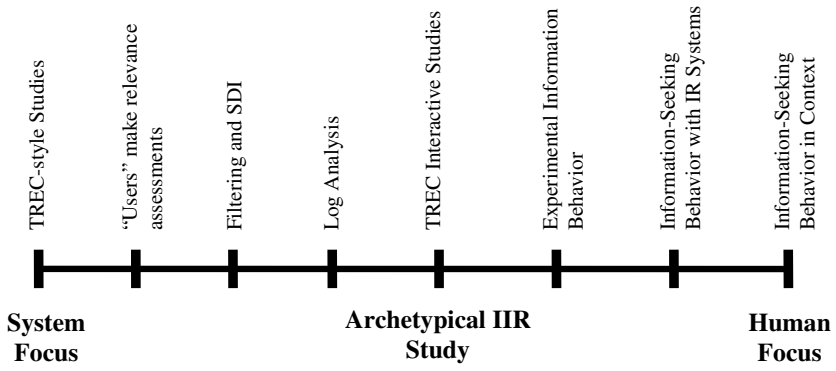

Fig. 2.1 Research continuum for conceptualizing IIR research.

search experiences and behaviors, and their interactions with systems. This type of study differs from a pure system-centered study because researchers recruit users to make assessments and build new infrastructure, rather than relying on the TREC infrastructure. This is often done because researchers are working on new problems or tasks that have not been addressed by TREC. For example, Teevan et al. [269] studied relevance feedback and personalization; this required the collection of queries, documents, and relevance assessments from users. Although it is possible to study the interaction between the user and the information need, or the user and documents, this is usually not the focus of this type of study.

Intent and purpose of the research are important in determining where a study belongs along the continuum. Consider a study where a system evaluation with users has been conducted but the researchers are primarily interested in demonstrating the goodness of the system, rather than understanding the user-system IR interaction; the user study is, in effect, an ancillary task rather than a central focus. In many ways, these types of studies undermine efforts to create a more solid foundation for IIR studies, since users are essentially treated as black boxes. Although it is not claimed that all IR studies should focus on users, an explicit mention of the focus of the study should be made so that readers can better distinguish between findings about IR systems, findings about interactive IR and findings about users. There is also

the question of whether it is even possible to claim that such a study is not at least in part a study of human behavior since users interact with the system and these interactions impact the system's responses. Although it is often claimed that the system is being evaluated and not the user, in practice this is difficult to do since a user is required for interaction.

As we move along the continuum, we come to studies where both the system and user are of interest, but there is still no interactive searching. These types of studies are less common today, but as the push towards personalization continues, we may see more studies of this type. A classic example of this type of study is evaluation of systems for the selective dissemination of information (SDI) from the 1970s, where the common evaluation model was to have users construct profiles and evaluate documents that were selected by the profiles [187]. Although users did not engage in searching, there was an attempt to understand how best to represent and update the user's information need and how best to present documents to users. The interaction, in these types of studies, was among the user, the information need and the documents. More current studies of proactive IR systems that observe users' behaviors while they do some activity (e.g., searching or writing) and fetch related articles are also in this class [41, 92].

Studies using transaction logs fall next on the spectrum. Although such studies have been around for quite some time, the availability of large search logs, in particular from search engine companies, have made this type of study very popular. Currently, the most popular type of analysis that is conducted with search engine logs looks at queries, search results and click-through behavior. These types of studies are primarily descriptive rather than explanatory, even though it is possible to model user behavior and interactions for certain situations. While many assumptions must be made about user intention, the shear amount of data that is available allows researchers to identify important regularities in Web search behavior. Furthermore, it is possible to manipulate some aspects of the search experience in a live test of a new interface feature or ranking algorithm and use the log data as a way to observe potential differences in performance [11]. Transaction log

analysis is often used in many types of studies, but studies represented at this place on the continuum use logs as the primary data source. Example studies are Agichtein et al. [3], Bilenko and White [28], and Jansen and Spink [146].

The next type of study is embodied by the TREC Interactive Track evaluation model. In this type of study, a system or interface feature is typically being evaluated. The design of such features is usually related directly to the human — whether it is human behavior and cognition, or information seeking context. The goals of such features are usually to better support the search process. This includes not only finding useful documents, but also supporting cognitive activities associated with search, such as sense-making. Studies located around this point on the continuum employ multiple methods of data collection, and usually include measures of system performance, interaction and usability. It is also common for studies around this point to use interviews to obtain qualitative feedback from users about their experiences. This point represents the classic or archetypical IIR study, which is the primary focus of this paper. Example studies are Joho and Jose [154] and White et al. [292].

As we move along the continuum, the next type of study focuses more on information behavior. In this type of study, the researcher controls aspects of the search process that are not typically controlled in the classic IIR study. For instance, the researcher might control what results are retrieved in response to a user's query or the order in which search results are presented to users. The point of such studies is to isolate and study individual aspects of the search process, rather than the entire process. One difficulty in studying the entire process is that each person experiences search differently; the goal of studies represented by this point on the continuum is to make these experiences as similar as possible so that causality can be studied with greater confidence. Studies represented by this point often use experimental methods that are commonly used in psychology. These studies focus on a slice of the search process and manipulation and control are often used. These studies are generally more interested in saying something specific about behavior, rather than on demonstrating the goodness of a particular IIR feature or system. Example studies are Arapakis and

Jose [12], Huang and Wang [135], Joachims et al. [152], Kelly et al. [168], Smith and Kantor [243], and Turpin and Scholer [276].

The next set of studies focus on general information search behavior in electronic environments. Most often there is no experimental system involved; researchers are instead interested in observing and documenting people's natural search behaviors and interactions. This might include studies of users' search tactics, studies of how users make relevance assessments, or studies of how users re-find information on the Web. Studies at this point also include those of searcher intermediaries and other professional searchers. Although these people are searching on behalf of others, they interact with the system to retrieve information. These studies differ from those mentioned above primarily in the methods they employ and the amount of the search process they examine. These studies often lead to better understandings of users' natural search behaviors, which is fundamental to the development of better IIR systems. However, these studies are not always driven by system concerns. Example studies are Byström [43], Ford et al. [99], Kellar et al. [163], and Kim and Allen [171].

Finally, the human-end of the spectrum can be characterized by research that focuses most exclusively on humans, their information needs and information behaviors. Researchers often insert themselves into a setting as an observer and gather data using qualitative techniques such as observation and interviews. In these studies, investigators explore the real information needs of users and their subsequent information seeking, within the particular context in which these needs arise without regard to a particular type of IR system. Solomon [244] provides a review of this type of research. Other work of this type can be found under the heading of everyday life information seeking [241]. Although results of these studies might inform the design of IR systems, this is usually not the primary goal. Example studies are Chatman [51] and Hirsh and Dinkelacker [131].

The primary focus of this paper is the archetypical IIR study, which represents IIR system evaluation with users. The archetypical IIR study represents a good entry point into IIR since it has a relatively long history, is most connected to traditional IR research, and is somewhat balanced with respect to computer and behavior sciences. Similar to

traditional IR evaluation, IIR evaluation has historically emphasized effectiveness, innovation, and application, sometimes at the cost of more basic scientific findings about interaction and behavior. Despite often being called *user-centered*, much of this research has primarily been about systems. As IIR evolves, it is hoped that equal emphasis will be placed on research that develops explanations of user behaviors and interactions that extend beyond an individual system.

Studies that span the spectrum from information seeking in electronic environments to log-based studies will also be discussed at some level because such studies form the core of IIR. Studies at either end of the spectrum will not be discussed since they are not considered as a core part of IIR, but rather as two different areas that frame IIR research. For a comprehensive discussion of theories of information seeking and methods for investigating information seeking behavior, readers are referred to Fisher et al. [93], Case [49] and Wildemuth [294]. For a discussion of common evaluation models used in systems-centered research, readers are referred to Spärck-Jones [246] and Robertson [220].

3

Background

3.1 Cognitive Viewpoint in IR

One important event which changed the way in which users were conceived and integrated into IR research was the *Workshop on the Cognitive Viewpoint* in Ghent, Belgium [70]. De Mey is noted as having first articulated the cognitive viewpoint at this workshop, although Ingwersen and Järvelin [139] note that this general viewpoint had been expressed by others at the time of De Mey's articulation. The cognitive viewpoint offered the first coherent alternative to what is known as the systems or algorithm viewpoint in IR [140]. The systems or algorithm viewpoint is the one most familiar to IR researchers — this viewpoint is embodied by the Cranfield and TREC evaluation models [288]. This viewpoint focuses on the system and makes a number of simplifying assumptions about users, and their needs and behaviors. Researchers adopting the system perspective are not blind to these assumptions, but maintain they are necessary to isolate and study individual aspects of the system. These assumptions have lead to the development of some strong evaluation norms for researchers working from the systems perspective.

The cognitive viewpoint embraces the complexity inherent in IR when users are involved and focuses attention on the cognitive activities that take place during information seeking and retrieval, and user-information, user-system interactions [138]. Ingwersen and Järvelin [139] identify five central and interrelated dimensions of the cognitive viewpoint[1]: (1) information processing takes place in senders *and* recipients of messages; (2) processing takes place at *different* levels; (3) during communication of information any actor is *influenced* by its past and present experiences (*time*) and its social, organizational and cultural environment; (4) individual actors *influence* the environment or domain; and (5) information is *situational* and *contextual*. While it is clear in viewing these dimensions that the cognitive viewpoint focuses on the user, Ingwersen and Järvelin [139] are carefully to point out that the cognitive viewpoint is not just about users' cognitive structures, but also about the numerous other cognitive structures represented in the IR system.[2] For instance, those represented by document authors and system developers. An examination of these five dimensions exemplifies some of the difficulties with conducting studies from this perspective. Specifically, that each individual user experiences IR in a different way, and that as soon as a user begins interacting with an IR system a series of cognitive changes take place which are unobservable, but will likely affect the user's subsequent interactions and behaviors. Similar to researchers working under the system perspective, researchers working under the cognitive perspective must also make simplifying assumptions and abstractions about some parts of the process.

Which perspective — the system or cognitive — is behind IIR research? Given the emphasis on the user in IIR it would appear that such research is guided by the cognitive perspective. However, a close look at discussions surrounding these two perspectives raises some questions. The systems or algorithm viewpoint is often referred to as the laboratory approach. If this is the case, then IIR surely falls under

[1] Emphasis is from Ingwersen and Järvelin [139].

[2] The cognitive viewpoint uses *system* in a more general sense to mean a combination of things or parts — this includes the technology, user, and environment.

the systems viewpoint since many IIR studies are in fact conducted in a laboratory.[3] Borlund [33] states that the systems viewpoint is concerned with achieving reliable results through control of experimental variables and repeatability of experiments, while the user-centered approach[4] is concerned with studying IR under real-life operational conditions. This would seem to preclude any controlled laboratory experiments from the cognitive viewpoint, another indication that perhaps IIR is part of the system perspective. Borlund [34] notes that a hybrid approach is needed for IIR that combines elements from the systems and cognitive perspectives and goes on to propose a framework for the evaluation of IIR systems. Ingwersen and Järvelin [139, p. 21] hint that IIR is a part of IR and represents the system perspective, perhaps because many IIR evaluations use standard TREC collections and thus, make simplifying assumptions about the nature of relevance. However, they describe IIR as being concerned with the "communication processes that occur during retrieval of information by involving all major participants in information seeking and retrieval, i.e., the searcher, the socio-organizational context, the IT setting, interface and information space". Although IIR evaluations have characteristics of each of the two major perspectives, fundamentally IIR is about humans, cognition, interactions, information and retrieval, and most IIR researchers would probably align their research with the cognitive perspective. Although many IIR evaluations use standard test collections and make simplifying assumptions about things such as relevance assessments, many of these studies are still being conducted to further our understanding of how people interact with systems to do IR.

3.2 Text Retrieval Conference

Three Tracks have attempted to develop a TREC-style evaluation framework for studying interaction and users: the Interactive Track (TRECs 3–11), the HARD Track (TRECs 12–14), and ciQA

[3] It is more likely the case that the label "laboratory" is inappropriate since it is a place, not an approach.

[4] According to Järvelin [147] the user-centered approach was subsumed by the cognitive perspective in the 1990s.

(TRECs 15–16). Each of these Tracks experimented with different types of evaluation frameworks, but none were successful at establishing a generic evaluation framework that allowed for valid cross-site comparisons. Of the three Tracks, the Interactive Track was the most responsible for developing some of the first accepted protocols and measures for IIR evaluation.

3.2.1 TREC Interactive Track

The TREC Interactive Track lasted nine years and made some of the most important contributions to the development of a method for IIR system evaluation [80, 128, 175]. The work of this Track will be summarized here; interested readers are referred to the chapter by Dumais and Belkin [80] in the edited book describing TREC [288].

In the initial year of the Track (TREC-3), participants were required to recruit subjects who were tasked with creating an optimal routing query. Fifty standard routing topics were provided and participants were allowed to recruit any number of subjects so long as each participating site contributed at least one routing query for each topic. Subjects developed their queries using training data and there was no standard protocol by which participating sites were required to operate. Participating sites could investigate any aspect of interactive behavior, including the influence of different system features on behavior. Participants experimented with a variety of interface and system features to assist subjects in completing the routing tasks. While subjects searched the training database when constructing their queries, they did not search for and save documents that they believed were relevant to the topic. Their only job was to construct routing queries. Major findings were that automatic techniques for creating routing queries performed better than human techniques, that the routing task was difficult for subjects to do and that a lack of standard protocol made it difficult to compare results.

In the next iteration of the Track (TREC-4), an *ad-hoc* search task was used instead of a routing task. The *ad-hoc* task required subjects to find and save as many relevant documents as possible for 25 topics. Subjects were also asked to create a final, best query for

the topic. Again, there was no standard protocol for administering the experiments, but participants were required to have at least one search conducted for each topic. Participants were also required to log each search, provide specific kinds of descriptive data about search behavior, including a narrative account of at least one search. Although subjects' job was to find and save documents they considered relevant to the topics, their judgments and performances were evaluated using relevance judgments made by TREC *assessors*, who created topics and established the benchmark relevance assessments.[5] Major findings from this iteration of the Track were that subjects' relevance assessments differed from assessors and that standard TREC metrics were not suitable to interactive searching scenarios because they were in part measuring other things besides performance — in particular, the extent to which subjects' relevance assessments matched assessors' — and they were computed based on 1000 retrieved documents which does not make much sense in an interactive setting since users are unlikely to examine 1000 documents.

The next year of the Track (TREC-5) saw a decline in participation, but ultimately laid the foundation for the next several iterations of the Track. In the previous two years the Track had four and ten participating sites, respectively, but only two sites completed participation in the Track at TREC-5, although others were involved with planning the Track [17, 21]. Based on experiences in TRECs 3 and 4, it was decided that the best approach would be to develop a method for comparing the performance of different IIR systems at different participating sites, rather than comparing human and system performance. For this Track, a new task was created which did not correspond to tasks used in other established TREC Tracks (e.g., routing and *ad-hoc*). This task was dubbed the *aspectual recall task* and required subjects to find documents that discussed different aspects of a topic rather than all documents relevant to a topic. This task was used by the Interactive Track in TRECs 5–8, with only slight modifications. Twelve topics

[5] Usually TREC assessors are recruited by the National Institute of Standards and Technology (NIST) which manages TREC. In most cases, these people are retried intelligence analysts, but some Tracks have used other kinds of people for topic creation and assessment.

were created based on previous *ad-hoc* topics (as compared to 50 routing topics in TREC-3 and 25 *ad-hoc* topics in TREC-4). Participating sites were required to provide the list of documents saved by subjects, search logs and a narrative account of one search. TREC assessors used the saved documents to compile a list of unique aspects and a key that showed which documents discussed which aspects. Aspectual recall was then computed, as well as standard precision.

In addition to these guidelines, the Track also developed an experimental design that required participating sites to implement two systems: a baseline, which was provided to participants, and an experimental system, which participants created. The Track also created topic and system rotations to control for order effects that required participating sites to study at least four subjects. These additional requirements were likely the reason why so few participants managed to complete the Track; it is reported in Dumais and Belkin [80] that even the two sites that did participate were unable to complete the experiment as planned. Thus, one of the major findings from this iteration related the amount of resources and time required to craft and implement the experimental design. There were concerns about the small number of topics (keeping in mind that topics are sampled just as users are) and results showed strong subject effects, topic effects and subject–topic interaction effects.

The TREC 6–8 Interactive Tracks basically used the same evaluation model that was developed in TREC-5 with some minor changes. Nine sites were able to complete the study in TREC-6 and it is described by Dumais and Belkin [80, p. 136] as being "the first true cross-site comparison in the interactive track". Only six topics were used, which allowed each subject to search for all topics (unfortunately this meant that each subject's experimental session lasted approximately 3 hours). The required, shared baseline system remained. The amount of data provided by the nine participating sites allowed for a more rigorous statistical analysis which found significant effects for topic, subject, and system. It was also the case that participants desired to create their own baseline systems that would allow them to better focus on interactive techniques that interested them. This major change happened in TREC-7. Note that this meant that cross-site comparisons

were no longer possible since each participating site created their own experimental and baseline systems.

TREC-9 represented a departure from the aspectual recall task which was in part motivated by a desire to reduce the amount of time required of subjects and explore additional types of tasks and collections. In this iteration, eight fact-finding tasks were created which consisted of four *n-answer* tasks (these tasks required subjects to find some number of answers in response to a question) and four *specific-comparison* tasks (these tasks required subjects to identify which of two provided answers was correct). The design required 16 subjects — a large increase over previous years, but since subjects were only given 5 min to answer each question, this resulted in experimental sessions that only lasted about one hour. Subjects were required to identify the answer and to save documents that helped them determine the answer. These answer-document pairs were then analyzed to determine performance.

The TREC-9 design only ran once and the Interactive Track moved to a 2-year cycle with TRECs 10–11 and focused on Web search. The goal of the first year was to define important issues and tasks worthy and possible of study in the Web environment. Subjects were provided with eight topics and searched the open Web. No standard instruments, experimental protocols or systems were required. In the second year, participants used the .gov collection from the TREC-11 Web Track and the Panoptic search engine was made available to participants [125]. The framework returned to a more tightly controlled experiment with a required design, protocol, and instruments.

Finally, in its last year (TREC-12), the Interactive Track was a subset of the Web Track. Subjects were asked to complete a topic distillation task, which asked them to construct a resource page (list of useful URLs) for a particular topic. The .gov corpus was used and two versions of the Panoptic search engine were made available to participants. Participants followed a specified experimental design and protocol. Interestingly, in this task, lists generated by human subjects were compared with lists generated automatically by systems, which was one of the original points of comparison in the Track during TRECs 3 and 4 (i.e., automatic versus human).

The work of the TREC Interactive Track resulted in many discoveries about IIR evaluation. First, assessors' relevance judgments were not generalizable and using these judgments to evaluate the performance of others was fraught with difficulty. Second, many standard evaluation metrics developed to assess system performance were not particularly useful in interactive settings. Third, including large numbers of topics in a laboratory evaluation was not reasonable given human limitations (both physical and cognitive). Larger numbers of topics require larger numbers of subjects and resources.

Although it was not possible to create an IIR evaluation model that was similar to the standard TREC evaluation model, the Interactive Track played an important role in establishing standards for IIR evaluation. This included a standard design and protocol, as well as standard techniques for reporting search logs and other data. The evaluation model of the archetypical IIR evaluation study derives directly from the work of this Track.

3.2.2 TREC HARD Track

The High Accuracy Retrieval of Documents (HARD) Track followed the Interactive Track and ran for three years [4, 5]. The set-up of this Track varied from year-to-year, but the primary focus was on single-cycle user-system interactions. These interactions were embodied in *clarification forms*. In most cases, participating sites used these forms to elicit feedback from assessors. Thus, the interaction consisted of a single exchange between the system and assessor. One of the initial goals of this Track was to represent and incorporate aspects of assessors' context into retrieval, thus, in addition to a corpus, topics and relevance judgments, this collection also contained user metadata describing context. The types of interactions that could occur were defined by the Track and there was no interactive searching performed by assessors. Instead, assessors completed these single-cycle interactions remotely (i.e., the assessors did not visit each participating sites' laboratories). In the first version of the Track, TREC assessors were also subjects — this was different from most iterations of the Interactive Track where assessors and subjects differed. In one iteration of

this Track, assessors represented a different kind of person than the traditional TREC assessor; they were interns and personnel at the Linguistic Data Consortium.

While the types of interactions were limited to those that could occur during a single exchange, the Track provided participants with an opportunity to engage in interactions with the assessors and elicit feedback from them. Common rules of the Track governed that some aspects of the interactions were the same across participating sites. To a certain extent assessors were also held constant since they each completed all the interaction forms for their particular topics. However, because the same assessor completed a number of interaction forms for a given topic, it was impossible to control for learning effects in this evaluation model — with each interaction assessors learned more and more about their topics, so while the Track studied single-cycle interactions, assessors actually engaged in multiple interactions (even if these were with different clarification forms). Although the evaluation model is not optimal for cross-site experimentation and full-scale IIR system evaluation, it can be used in a single-site study where researchers are interested in isolating and studying individual aspects of the search process while holding other aspects constant (e.g., [164]).

3.2.3 TREC ciQA Task

Following the HARD Track, ciQA (complex interactive question–answering) was introduced as a sub-task of the QA Track in 2006 and 2007 [68, 167]. The first year of this task was modeled closely after the HARD design where assessors completed forms that had been created and submitted by task participants. The ciQA task differed from the HARD task in that the focus was on complex question answering, rather than the traditional *ad-hoc* document retrieval task. The task also did not attempt to incorporate context into retrieval. At the time, there was little research devoted to interactive QA of the type represented by the TREC QA Track and one goal of ciQA was to encourage exploration in this area.

The 2007 version of ciQA allowed for any type of interaction, including full-scale interactions with working systems. Assessors interacted

with experimental systems remotely via the Internet. However, this new setup did not eliminate the problem of learning effects since assessors still engaged in a number of interactions with a number of systems for the same topic. More than anything, ciQA represented a first attempt at studying interactive QA and deploying a large scale evaluation exercise remotely. One of the more interesting findings of ciQA was the extent to which assessors could be considered as regular system users.

4

Approaches

This section provides an overview of different research approaches used for evaluation. Several approaches are discussed, but the emphasis in this paper is on laboratory evaluations.

4.1 Exploratory, Descriptive and Explanatory Studies

One way to think about research approaches is to consider specific goals of the research: exploration, description or explanation. Such characterizations can be found in almost any research methods textbook (e.g., [13]), but are useful to consider here since they suggest how studies should be evaluated. If the study goal is description, then evaluation criteria for explanatory studies should not be applied during the review.

Exploratory studies are typically conducted when little is known about a particular phenomenon. Exploratory studies often employ a variety of research methods with the goal of learning more about a phenomenon, rather than making specific predictions. Exploratory studies often have less structured methods than descriptive or explanatory studies and it is often the case that results from exploratory studies lead to descriptive or explanatory studies. Research questions are typically broad and open-ended and hypotheses are uncommon.

Descriptive studies are focused on documenting and describing a particular phenomenon. Examples of descriptive studies are those whose results characterize query logs and query behaviors (e.g., [146]). The main purpose of such studies is to provide benchmark descriptions and classifications. Although results of descriptive studies can become dated over time, temporal comparisons of results can made. As with exploratory studies, results of descriptive studies can be used to inform other studies. For instance, an analysis of a query log could give a researcher a principled way of selecting tasks to use in a laboratory study or suggest a hypothesis that can be evaluated as part of an explanatory study. Descriptive studies can lead to a weaker form of prediction via correlation analysis. However, such studies are not able to explain why a relationship exists between two variables.

Explanatory studies examine the relationship between two or more variables with the goal of prediction and explanation. Explanatory studies are often concerned with establishing causality and because of this require variables of interest to be isolated and studied systematically. Explanatory studies are often conducted in the laboratory since this is the environment that affords the researcher the most control over the situation. Explanatory studies use more structured and focused methods than exploratory or descriptive studies and involve hypothesis testing. Despite the name, it is important to note that not all explanatory studies offer explanations — many just report observations and statistics without offering any explanation. It is also important to distinguish between prediction and explanation: it is possible to build predictive models of events without actually understanding anything about why such events occur. Very often researchers stop at prediction and do not pursue explanation, but it is actually explanation that is tied most closely to theoretical development.

4.2 Evaluations and Experiments

In classic IR, experiment and evaluation have been used interchangeably, but these two types of studies need to be separated when discussing IIR. One can conduct an evaluation without conducting an experiment and *vice versa*. Evaluations are conducted to assess the

goodness[1] of a system, interface or interaction technique and can take many forms (some of which are discussed later). Experiments have historically been the main method for interactive system evaluation, but experiments can also be conducted to understand behavior. IIR experiments look similar to those conducted in social science disciplines such as psychology and education. For instance, it is common to evaluate the relationship between two or more systems and some set of outcome measures, such as performance or usability. This is a standard experimental model where the goal is to examine the effects of an independent variable (e.g., system type) on one or more dependent variables (e.g., performance and usability). Two important characteristics of experiments are that there are at least two things being compared (e.g., system type) and that some manipulation takes place. For instance, one might manipulate which system a subject uses.

In some types of IIR studies only a single system is evaluated. This is a weaker form of evaluation since it is not possible to demonstrate how much better users perform or how different their behaviors and interactions are since there is no point of comparison. Traditional usability tests are examples of this type of evaluation. Traditional usability tests are usually conducted with a single version of a system, with the goal of identifying potential usability problems. These types of studies are particularly important for *formative evaluation*: evaluation that is conducted during system development [97]. Formative evaluations can be contrasted with *summative evaluations* which assess the value of a finished or mature system.

4.3 Laboratory and Naturalistic Studies

Studies can also be characterized according to where they take place. Studies can take place in the laboratory or in a naturalistic setting. Most, but not all, experiments take place in the laboratory. It is important to note that if you are conducting a study in a lab, this does not automatically make the study an experiment. It is traditional to conduct other types of studies, such as usability tests, in labs even

[1] The term *goodness* is used as an abstract construct and may, of course, represent a number of things such as performance, usability or effectiveness.

when there are no experimental conditions or manipulations. Laboratory studies are good with respect to the amount of control researchers have over the study situation. This is particularly useful when trying to isolate the impact of one or more variables. Of course, one perennial criticism of laboratory studies is that they are too artificial, do not represent real life and have limited generalizability.

Naturalistic studies examine IIR in the settings in which it occurs. Log-based studies are examples of naturalistic studies since they capture behavior as it is occurs in real life. The behavior that is captured is thought to be more representative of the user's true behavior since the chances that it contaminated or biased by the study design or the researcher is much less than that captured in a laboratory. One important drawback to conducting naturalistic studies is that the researcher has little control over the setting, which can make it hard to make cross-user comparisons. The amount and types of data one collects might be as variable as the number of users in the study. It can also be difficult to administer naturalistic studies since they are more intrusive and the user often has to be willing to give up some privacy.

It is also possible to conduct natural experiments. One example is the study by Anick [11] who conducted live trials of an interface for query expansion. As a researcher at a large search engine company, Anick was able to distribute an experimental interface to a number of users and compare its use to the standard interface. In another example, Dumais et al. [79] deployed two working versions of a desktop search tool to 234 people within an organization and gathered data on its use over the course of six weeks.

4.4 Longitudinal Studies

Naturalistic studies are often longitudinal in that they take place over an extended period of time and measurements are taken at fixed intervals. Longitudinal approaches can also be incorporated into laboratory studies — users might be required to attend multiple sessions over time. Longitudinal approaches are often used when one wants to study if and how something changes over time. Although studies employing longitudinal approaches are more time consuming, they represent a necessary

and important type of study since many kinds of information seeking and retrieval activities take place over extended periods of time during multiple search sessions [181, 247]. Only studying single search sessions, such as the kind typically studied in the lab, places limits on what is known about IIR.

One important consideration in designing longitudinal studies is determining the duration of the study, as well as the measurement interval. Having some understanding of the behavior and some expectation about how often it occurs can help one make this decision. For instance, if the behavior occurs everyday then one might want a different duration and measurement interval than if the behavior occurs weekly. It is also the case that in the social world, user behavior can be governed by a number of external factors. For instance, the occurrence of a holiday or a project deadline will likely change the kinds of behaviors users exhibit and these behaviors may not represent their typical behaviors.

4.5 Case Studies

Another approach is the case study. This type of study is not seen a lot in IIR, but as the area grows, researchers may begin to do more of these studies. Case studies typically consist of the *intensive* study of a small number of cases. A case may be a user, a system or an organization. Case studies usually take place in naturalistic settings and involve some longitudinal elements. Researchers conducting case studies are less interested in generalizing their research results and more interested in gaining an in-depth, holistic, and detailed view of a particular case. The ability to generalize is traded for a more complete and robust representation of how something occurs in a small number of cases. Case studies are particularly useful when little is known about an area and for understanding more about details that sometimes get lost when averaging over large numbers of users.

4.6 Wizard of Oz Studies and Simulations

Wizard of Oz studies get their name from the well-known film/book of the same title. In this work, the protagonist Dorothy, travels a

great distance to visit the Wizard of Oz who at first glance appears very grand and intimidating. However, as it turns out, the Wizard is really just a small man behind a curtain orchestrating a grand Wizard façade. Wizard of Oz studies are similar in that researchers often imitate 'grand' systems that they would like to study. While users believe they are interacting with a real system, in reality there are one or more researchers 'behind the curtain' making things work. For example, suppose a researcher wanted to study a speech user interface for querying and interacting with an IR system. Rather than building the entire system, the researcher might first want to learn something about the range of desired communications and interactions. Users might be instructed to speak to the system while a researcher sits in another room and controls the system. Wizard of Oz studies can be used for proof-of-concept and to provide an indication of what might happen in ideal circumstances [67].

Wizard of Oz studies are simulations. Heine [126] discusses simulation experiments in the context of IR research. While systems are simulated in Wizard of Oz studies, another entity that has been simulated in IR studies is users. Rather than recruit and study real users, simulated users are used to exercise systems. Users may be defined by one or more characteristics which can take on a number of values. Simulated users consist of various combinations of these characteristics and values. Simulated users can also represent different actions or steps a real user might take while interacting with an IR system. Some concerns about the use of simulated users include the realism and utility of the simulated users and more fundamental issues about what it means to really study users and user behavior. Some positive things about simulated users is that IIR evaluations can be conducted more rapidly and with less cost and larger numbers of users (albeit simulated), who have a broad range of carefully controlled characteristics. It is also the case that one can control learning through programming — the researcher is able to determine if and how learning will take place during the study. Example studies of simulated users include Lin and Smucker [180] and White et al. [293].

5

Research Basics

This section discusses some of the basics of IIR research. Most of these things are necessary parts of any empirical research, but they are reviewed here because they form the foundation of research endeavors.

5.1 Problems and Questions

All studies are motivated by some problem or gap that exists in the research. Thus, the first step in conducting any study is to identify and describe the problem. This helps focus one's attention and provides a roadmap for the presentation of research results. Describing the problem and its relevant pieces also helps ensure that proper attention has been paid to what is currently known about a particular issue (and that a proper literature review has been conducted). Within the context of some given problem, the research question essentially identifies the piece of the problem that will be addressed by the study.

The research question[1] should be narrow and specific enough that it can be addressed in a study, but the specificity of the question will

[1] Although research question is used in the singular, it is common for studies to have multiple related research questions. It is advised to have multiple, simple questions rather than a single, complex question.

depend in part on the purpose of the study. For instance, explanatory studies typically have much more concise and narrow research questions than exploratory studies. Research questions should also be value-free in a sense that the researcher's opinion is not embedded in the question. If the researcher has some belief about the outcome of the study, then this should be framed as part of the study hypothesis, not research question.

Figure 5.1 shows some example research questions from IIR studies. The first example identifies a broad question that was part of an exploratory study. Example 2 is a descriptive question, while the final two examples are of specific, focused explanatory research questions. Since description often leads to explanation, some studies might have both a descriptive and explanatory research question. What is known about a particular phenomenon and the extent to which one wants to study it determines the specificity with which questions can be asked.

Researchers who are very focused on their particular system might also be tempted to pose a question in the form, "Is System X better than System Y?" This is probably okay (even though it could technically be answered with a binary response), but in some ways detracts from more specific types of questions that can be asked regarding the differences between System X and System Y. A better approach would be to identify the expected differences (e.g., in performance or usability) and formulate specific questions about these differences, rather than to lump everything together in a single question. The general question might be fine for a strict evaluation, but more specific questions are

Example 1: How do people re-find information on the Web? [268]
Example 2: What Web browser functionalities are currently being used during web-based information-seeking tasks? [163]
Example 3: What are the differences between written and spoken queries in terms of their retrieval characteristics and performance outcomes? [62]
Example 4: What is the relationship between query box size and query length? What is the relationship between query length and performance? [22, 159]

Fig. 5.1 Some example research questions from IIR studies. Example 1 is exploratory, Example 2 is descriptive, and Examples 3 and 4 are explanatory.

better suited for situations where the researcher wants to understand IIR behaviors or phenomenon that transcend a specific system.

5.2 Theory

A theory is "a system of logical principles that attempts to explain relations among natural, observable phenomena. A theory appears in abstract, general terms and generates more specific hypotheses (testable propositions)" [95, p. 132]. Littlejohn [182, p. 21] describes theory as "any conceptual representation or explanation of a phenomenon." Thus, two of the most important qualities of theories are that they offer explanations of particular phenomena and allow researchers to generate hypotheses. The testing of hypotheses, in turn, allows researchers to further refine and extend theories. Theories can also be developed and refined through grounded theoretical approaches [108].

Historically, IR and IIR have been driven by innovation and technology; the emphasis has been on applied and practical aspects of science. As a result, there has been less in the way of theoretical development. Indeed, a lot of research does not even mention or consider theory. This is not to claim that there are no theories or theoretical constructs in IR and IIR — some examples include Robertson's [218] probability ranking principle, Ingwersen's [137] theory of polyrepresentation and Belkin's [19] anomalous states of knowledge. In the area of human information behavior there is even a book describing common theories and models [93]. What is claimed is that at present, theory innovation has received less attention than system innovation. Currently, research in IR and IIR emphasizes results over explanation, and many studies are not motivated theoretically.

5.3 Hypotheses

Hypotheses follow from research questions (or theory) and state expected relationships between the concepts identified in the questions (such concepts may be more or less definable, but they are eventually represented by variables). There are two types of hypotheses: *alternative* hypotheses and *null* hypotheses. An alternative hypothesis is

the researcher's statement about the expected relationship between the concepts under study. This is also known as the research hypothesis. Research hypotheses are called *alternative* because they present alternatives to the null hypothesis, which states that there is no relationship or difference. The null hypothesis is accepted by default; the scientific method places the burden on the researcher to demonstrate that a relationship exists. Although there are several published accounts of researchers proposing and testing null hypotheses, this is actually counter to the scientific method since null hypotheses do not need testing — they represent the default description of things. Scientists start with null hypotheses because logically it is easier to show that something is false instead of true.

In general, science is about accumulating evidence to demonstrate some relationship rather than providing definitive proof. The logic is such that we are not ever able to say that our alternative hypothesis represents the *truth* or that we have *proved* it. In fact, it can be argued that it is not useful to talk about truths, especially when studying the social world, but rather to talk about accumulation of evidence, which supports a particular hypothesis or points in a general direction. When we engage in hypothesis testing, strictly speaking we are able to make two statements about the relationship between the evidence we collect and our hypothesis: (1) our evidence allows us to *reject the null hypothesis*, in which case it is shown that our hypothesis provides a better (but not the only) description of what is going on, or (2) our evidence does not allow us to reject the null hypothesis, in which case *we fail to reject the null*. The burden of proof lies with the researcher, and even then, absolute proof is not a useful construct. Instead, researchers show that the evidence they collect demonstrates that the null hypothesis does not adequately describe what is going on; the alternative hypothesis offers an alternative explanation of what is going on, but it still may not adequately capture what is happening.

At the most basic level, a hypothesis should state a relationship between two or more things. One common mistake that many researchers make is not actually posing testable hypotheses or not fully articulating the comparison they would like to make. By nature, hypotheses are comparative and suggest the existence of at least two

things. For instance, it is not sufficient to say, "System A is usable," or that "System A is more usable," but "System A is more usable than System B." While one could provide descriptive statistics that show, for instance, that 70% of subjects rate System A as usable, without at least two groups, one cannot perform any statistical testing. There is no way to test the first statement, particularly since there are no benchmarks upon which to base rejection of the null. In the statistics section of this paper, it will be shown that it is possible to have a hypothesis that does not explicitly name two groups, but this is when one group is the population and the population parameter is known.

Hypotheses can also be directional or non-directional. The hypothesis that *System A is more usable than System B* is a directional hypothesis since the direction of the difference is given. Contrast this with a non-directional form of this hypothesis, *there is a difference in usability between System A and System B*. Finally, strictly speaking, hypotheses should be stated at or near the beginning of a study. Researchers often create hypotheses after they begin examining their data, but the scientific method calls for hypotheses to be identified clearly before any data are collected.

5.4 Variables and Measurement

Variables are present in almost all studies, although they play less of a role in qualitative studies. Variables represent concepts. Specifically they represent ways of defining, observing and measuring the concepts that researchers aim to study. Relevance, performance, and satisfaction are all concepts. To investigate concepts, researchers must engage in two basic processes: conceptualization and operationalization. These processes involve articulating definitions, but at two different levels.

5.4.1 Conceptualization and Operationalization

Conceptualization is the process by which researchers specify what they mean by particular terms. Some terms have very agreed-upon meanings. For instance, if we talk about someone's sex, most people would understand what we mean. However, other terms can have a variety of meanings. For instance, there is no universally agreed-upon

definition of relevance. We know from studies that have attempted to define relevance that there are many interpretations of this term, as well as many manifestations [235, 236]. Thus, the first step in attempting to measure a concept like relevance is to define it. Such a definition is considered a *working* or *nominal* definition — it represents a *temporary* commitment on the part of the researcher and helps frame the study and delineate findings. No claim is made about the universality of the definition.

Sometimes it is useful to subdivide a concept into dimensions to make conceptualization easier. For instance, in defining relevance, one might first identify specific dimensions such as those articulated by Saracevic [234] — algorithmic, topical, cognitive, situational, and affective. Next, one might provide specific definitions for these terms rather than trying to provide a single, all-encompassing definition for relevance.

After articulating conceptual definitions, the next step is to provide *operational* definitions, which state the precise way the concept will be measured. For instance, one might decide to measure topical relevance by asking subjects to indicate how useful they find documents and giving them a five-point scale to indicate this. One might define algorithmic relevance as the system's estimate of the likelihood that a user will find a document useful given a particular query and use the relevance scores produced by the retrieval system as an indicator. Usability, a concept that plays an important role in many IIR evaluations is often subdivided into dimensions such as effectiveness, efficiency, and satisfaction. Doing this is only part of the process since one must also define these concepts and state how they will be measured and observed. Very often researchers do not carefully articulate the conceptual and operational definitions that they employ in their reports; if this has not been done, it is difficult to evaluate the quality and appropriateness of the measures, and the validity of the work. This also makes it difficult to compare findings across studies.

5.4.2 Direct and Indirect Observables

A useful distinction to make between IIR measures relates to whether they are directly or indirectly observed. Direct observables are often

byproducts of a user's behaviors and interactions and are produced as the user searches. For instance, number of queries entered, number of documents opened, and the amount of time spent searching are examples of measures that are directly observable. Indirect observables are those things which cannot be observed and that essentially exist within the user's head. An example of an indirect observable is satisfaction.

For both types of observables — direct and indirect — it is important to ensure that the equipment used to make the observations is valid and reliable. Direct and indirect observables present different issues in this regard. With direct observables there is a ground-truth — we can physically count the number of queries a user enters and compare this value to what is recorded by some other instrument, such as a logger. However, for concepts that are only indirectly observable, instrumentation is more difficult. Not only must researchers be concerned with whether indirect measures are good representations of particular concepts, but they must also be concerned with how this information is captured (e.g., does a five-point Likert-type item adequately capture satisfaction?). Ground-truth exists in each individual user's head and there is no way to compare what we can observe through a self-report scale to this truth.

IIR is concerned with many more indirect observables than direct observables, which would suggest that measurement and instrumentation should be priority research issues. Unfortunately, there are not a lot of research programs focused on measurement, which makes it difficult to understand the extent of measurement problems in IIR evaluations. Many measures are developed in an *ad-hoc* fashion and there are few well-established measures and instruments, especially for indirect observables. Ultimately, any new measure should be both valid and reliable. These issues are discussed later.

5.4.3 Independent, Dependent, and Confounding Variables

Another distinction that can be made is between *independent*, *quasi-independent*, and *dependent* variables. Using the language of cause and effect, independent variables are the causes and dependent variables are the effects. In experiments, researchers typically manipulate the

independent variable — for instance, asking people to use particular systems or assigning them to particular experimental conditions. Quasi-independent variables are variables that can create differences in some outcome measure, but are not manipulated by the researcher. Sex is a good example of a quasi-independent variable. A researcher might be interested in examining differences in how males and females use an experimental and baseline IIR system, but a researcher cannot manipulate anyone's sex. The researcher might ask equal numbers of males and females to use each system, but this variable is not under the researcher's control in the same way as system type. Dependent variables are outcome variables, such as performance and satisfaction. In most IIR evaluations, researchers are generally interested in examining how differences in one or more independent variables impact one or more dependent variables.

Confounding variables (or confounds) are variables that affect the independent or dependent variable, but have not been controlled by the researcher. Often researchers are unaware that such variables exist until they begin a study or after a study ends. If a researcher realizes that such variables exist before the study starts, then the researcher can control the effects of the variables. For instance, a researcher might believe that search experience impacts how successful a person will be with an information retrieval system. If the researcher were testing two IIR systems, it would be important to ensure that equal numbers of subjects with high and low search experience were assigned to use each of the systems. If more subjects with high search experience were assigned to use one of the systems, then it might be found that subjects did better with this system, but the cause could not be attributed to the system since another potential explanation exists. In this case, search experience would be considered a confounding variable.

In addition to independent and dependent variables, moderating and intervening variables are also used to represent relationships among concepts. However, at present, these are used less frequently in IIR evaluations. Moderating variables affect the direction or strength of the relationship between an independent and dependent variable. For instance, consider the above example regarding the relationship between system, search experience and performance. Suppose the

researcher designed two interfaces, one which supported advanced search and another which supported simple search. The researcher might be interested in investigating how well subjects perform with each of the systems. The researcher might further believe that subjects who have high search experience will perform better with the advanced system, while those with low search experience will perform better with the simple system. In this situation, search experience is said to moderate the relationship between interface type and performance because different levels of search experience (high and low) result in different types of relationships between the independent and dependent variables.

Finally, an intervening variable provides a connection or link between an independent and dependent variable. For instance, a larger query box might lead subjects to enter longer queries. These longer queries, in turn, might lead to better performance. The connection between the independent and intervening variables and intervening and dependent variables basically represents two causal relationships. One would not say that a larger query box caused differences in performance, but rather that a larger query box caused differences in query length which in turn, caused differences in performance.

5.5 Measurement Considerations

When designing measures, there are a number of properties to consider. Most of these properties are related to measures that have a response set. Such response sets might contain numeric or textual choices or categories for responding.

The first property is related to the *range of variation* that is expected to occur in the concept being measured. The range of variation is the extent to which a measure presents an adequate number of categories with which to respond. Range of variation is closely related to the preciseness of the measure. For instance, when creating an instrument for eliciting relevance judgments from subjects, is a binary scale, a tertiary scale, or a five-point scale provided?

Exhaustiveness is closely related to range of variation and is the extent to which a response set can be used to characterize all elements

under study, where an element might be a system, document or user. For instance, with a binary relevance scale, a user might have a difficult time characterizing a document that is partially relevant. Another common example, although not used in most IIR evaluations, is race. Very often some races are missing from the list of choices (exhaustiveness) and others are not differentiated enough (range of variation). Understanding both the range of variation and how exhaustive measures need to be is very much related to the researcher's knowledge and expectations of the typical variation that exists within the elements under study.

Another property to consider is *exclusiveness*. This property is the extent to which items in the response set overlap. When this property has been violated, there might be more than one response that can be used to characterize a single object. For instance, a user might be provided with the following options for indicating relevance: not relevant, partially relevant, somewhat relevant and relevant. Most subjects would have a difficult time distinguishing between the middle two options (unless the researcher provided some very good definitions of each choice). Not all measures have to be exclusive, for instance, some questions allow a user to make more than one choice, but such questions can technically be viewed as a series of binary items.

Another property is *equivalence*. This property is the extent to which the items in a response set are of the same type and at the same level of specificity. Consider a scale that is meant to assess a person's familiarity with a search topic and has at one end of the scale the label *very unfamiliar* and at the other end, *I know details*. It would better to associate the first label with *very familiar*, and the second label with *I know nothing* since these are true opposites and at the same level (*knowing details* is slightly more specific than *being familiar*).

A final property is *appropriateness*. This is the extent to which the provided response set makes sense in relation to the question being asked. Consider the following question which might be asked of subjects, "How likely are you to recommend this system to others?" If the researcher provided subjects with a five-point scale with strongly agree and strongly disagree as anchors, then this response set would be inappropriate because the scale anchors do not match the question.

5.6 Levels of Measurement

The level of measurement used to represent a variable is a critical concept since it ultimately determines what types of statistical tests are possible. Researchers often wonder what type of statistical test is most appropriate for their data. The answer to this question, in part, lies with the levels of measurement. There are two basic levels of measurement: *discrete* and *continuous*. One of the biggest differences between discrete and continuous measures is the extent to which the values represent real numbers. This, in turn, impacts the extent to which data produced from these measures can be used in mathematical functions. As one moves from discrete to continuous levels of measurement, one is able to conduct more sophisticated types of statistical analyses.

5.6.1 Discrete Measures

Discrete measures provide and elicit categorical responses. These categorical responses can be textual or numeric. Discrete measures are divided into *nominal* and *ordinal* data types. Nominal data types provide response choices that represent different kinds of things but not different degrees. Ordinal measures provide response choices that are ordered, where choices represent different degrees. A classic example of a nominal measure is *sex*, which has two *levels* (choices or responses): male and female. These two levels are different from one another, but there is no order among them — one is not better or more than the other. Instead they represent the exhaustive set of choices that all subjects would need to be able to classify themselves. The most common type of nominal variable in IIR evaluations are independent variables such as interface type and task-type.

The two common ways that ordinal measures are used are as rank-order measures and as Likert-type[2] scale measures. An example: a rank-order measure is when a subject is given a set of documents and asked to order them from most relevant to least

[2] Likert-type is used to describe numeric scales (regardless of points) that are used to elicit data from users. The term *Likert-type* is used because the original Likert scale was a five-point scale that measured agreement and was scored and administered a bit differently than how such scales are often used today [178].

relevant. This type of measure allows the user to show the perceived relationship among the documents with respect to relevance, but this is a relative measure rather than absolute. For instance, we could identify which documents were more relevant than others, but we could not discuss the magnitude of these differences. That is, we could not say how much more relevant one document was than another. The difference between documents ranked 1 and 2 might be slight, while the difference between documents ranked 2 and 3 might be large. The differences between consecutive points are not equal.

The other common type of ordinal measure in IIR is Likert-type scales. While such scales give the impression of a true number line, the values do not represent real numbers (instead, one can think of the numeric values as labels). If we provided subjects with a five-point scale, where 1 = not relevant and 5 = relevant, and asked them to judge a set of documents, then we would still be in a position to describe which documents were more relevant than others, but we would be unable to describe the amount of these differences. We could say that a document rated 4 was more relevant than a document rated 2, but we could not say that a document rated 4 was *twice* as relevant as a document rated 2 since the scale contains no true zero.

To use a Likert-type scale subjects have to perform some calibration. This is unlikely to be consistent across subjects or even within a subject: one subject's 2 may not represent the same thing internally as another subject's 2. Because Likert-type scales are not represented by real numbers, the types of analyses that can be done with them are limited. This is unfortunate since these measures are the *sine qua non* of any science whose aim is to study human behavior and attitudes. Because of this, an accepted practice in the social sciences is to promote Likert-type measures to a continuous data type so that more sophisticated analyses can be done.

5.6.2 Continuous Measures

Continuous measures are divided into *interval* and *ratio* data types. For each of these types, differences between consecutive points are equal, but there is no true zero for interval scales. The most common examples

given for *interval* level data are the Fahrenheit temperature scale and intelligence quotient (IQ) test scores. For both measures, a score of zero does not indicate the complete absence of heat (indeed, the freezing point is 32 degrees) or intelligence. However, it is the case that the absolute differences between temperatures of 40 and 80 degrees, and 50 and 90 degrees are the same. However, it is not appropriate to say that a temperature of 80 degrees is twice as warm as a temperature of 40 degrees since there is no true zero on the scale.

The *ratio* level of measurement represents the highest level of measurement. A true number line underlies measures of this type. Common examples of ratio levels of measurement include time and almost any measure that can be verbally described as *the number of occurrences*. For instance, the number of queries issued, the number of pages viewed, and the number of documents saved. It is sometimes difficult to imagine these values being zero, but it is possible. For example, it is *possible* for someone not to enter a single manual query (imagine a system that supports browsing) or not open any documents during a search session. Another nice thing about ratio level data is that it can be transformed into ordinal and nominal level data. For instance, based on recall scores, an ordinal measure called *search expertise*, with the following levels: poor performers, medium performers and top performers could be created.

6

Experimental Design

The basic experimental design in IIR evaluation examines the relationship between two or more systems or interfaces (independent variable) on some set of outcome measures (dependent variables). IIR evaluations can include other independent variables as well such as task-type, and quasi-independent variables such as sex and search experience. One important part of experimental design which will be discussed in detail is rotation and counterbalancing. Tague-Sutcliffe [261] was one of the first to write formally about this in IIR. This allows one to control aspects of the study that might otherwise introduce experimental confounds. This section also presents other issues related to experimental design including study mode, protocols, tutorials, timing and fatigue, and pilot testing.

6.1 Traditional Designs and the IIR Design

Traditional designs can be discussed in terms of pre-experimental designs and experimental designs. These are standard designs that are discussed and presented in a number of research methods textbooks (e.g., [13]). They are not a creation of IIR and do not always fit perfectly with IIR study situations, but they do provide different

Group	Time_1		Time_2
1	O	E	O
2	O	C	O
3		E	O
4		C	O

Fig. 6.1 Solomon four-group experimental design.

ways of thinking about study design and measurement. The distinction between pre-experimental and experimental designs rests on the absence of a control group and baseline measurement. Figure 6.1 presents a well-known experimental design, the Solomon four-group design [47]. The different groups in this design can be used to illustrate other types of research design, including pre-experimental designs. A pre-experimental design with no control group or baseline measurement is represented by Group 3. In this group, an experimental stimulus (E) (e.g., a system) is introduced and then an observation or measurement (O) is taken of some outcome measure (e.g., performance). One of the most common types of studies in IIR that follow the design depicted by Group 3 is the single system usability test. There is no comparison or control system. Instead, subjects use one system and some initial feedback is collected regarding its goodness. Note that this type of study does not allow for the testing of hypotheses related to the system because there is only one system being studied. No comparison is possible, except with pre-determined population parameters, which are unlikely to exist. It is important that one looks closely at one system studies before deciding they are usability studies and not experiments. Many experiments only involve a single system, but some other variable of interest is manipulated and of interest. It is possible for IIR evaluations to have independent variables that are not tied directly to a system. The system may just be used as an instrument to facilitate information search (e.g., [168]).

The other attribute that makes this (Group 3 only) a pre-experimental design is that a baseline measure of the outcome variable has not been taken. In traditional experimental models, baseline measures of the outcome variables of interest are elicited before the stimulus is introduced. This is depicted by Group 1 in Figure 6.1. For instance, if one were evaluating a new drug designed to help people lose

weight and the outcome measure was a person's weight, one would need to obtain a baseline weight for each subject to know if the drug was associated with a decrease in weight — without this measure it would not be possible to determine this. In the context of IIR, one general goal of many evaluations is to determine if a particular system helps subjects find relevant documents. Attempting to elicit a baseline measure before the system (stimulus) is introduced does not make much sense and would probably not be possible. We can also imagine that the goal of an IIR system is to help subjects learn something about their information problems. To evaluate this, we would really need to measure subjects' knowledge of their information problems before and after they used the system.

Baselines are used in IIR evaluations, but in a way that differs slightly from the classic experimental model. In the context of IIR evaluations, baselines are often introduced as an alternative to the experimental system. Instead of taking a baseline measure before a user interacts with a stimulus, the baseline is more often represented by one level of the stimulus variable. For example, if the stimulus variable is an IIR system, it might have two levels: experimental and baseline. Thus, baselines in IIR evaluations are more similar to control groups (C) in Figure 6.1. In the *traditional experimental model*, the stimulus variable is usually either present or absent and a control group is used along with pre-treatment measurement (Figure 6.1, Groups 1 and 2). In Figure 6.1, the *classic IIR* design is represented by Groups 3 and 4. This model ostensibly functions as the archetypical IIR evaluation design.

A baseline (or control in the traditional model) is generally defined as the status quo, which raises some interesting questions with respect to IIR evaluations. Specifically, if IIR systems are under study and baselines represent subjects' normal experiences, then in most cases this would be a commercial search engine. However, it is not possible or valid to compare an experimental IIR system to a commercial search engine.[1] For instance, a researcher may be using a closed collection of newspaper articles; if a commercial search appliance were used to access

[1] This may be possible if you work for a commercial search engine company.

this collection as a baseline, it might not work optimally because of characteristics of the corpus and search algorithm. Thus, such an evaluation would not be comparing similar situations. Of course, whether a commercial search engine is a valid baseline depends greatly on the purposes of the study and the system. Even though it may not be possible or desirable to use a commercial search engine as a baseline, it is important to recognize subjects' previous search experiences and search norms will impact their interactions with, and expectations of, any experimental IIR system.

Developing a valid baseline in IIR evaluations often involves identifying and blending the status quo and the experimental system. For instance, if a researcher developed a new technique for displaying search results, then a baseline method of doing this could be modeled after methods used by commercial search engines. If the experiment was done using a proprietary system or well-established system, then the baseline could be the retrieval method currently used by that system (given that one was testing the workings of the system). Things get a bit more difficult when the experimental system or interface is something that subjects have never seen. Researchers often develop experimental IIR systems from scratch using languages such as Java. There is a good chance that the interface will look very different from a Web-based system to which subjects are accustomed. In this case, if one were comparing a new search interface feature, it would not be reasonable to compare this to a standard Web search engine since the number of differences between these two systems would be great. If differences were found, it would be difficult to relate them to the specific search interface feature of interest and to rule out the possibility that these differences were not caused by some other feature or aspect of the system.

As mentioned earlier, the design depicted in Figure 6.1 is called the *Solomon Four-Group Design* [47]. It was developed to address several major threats to the internal validity of experiments. These will not be discussed here, but suffice to say the four groups allow the researcher to control a number of threats to validity. The Solomon Four-Group Design is quite nice, but requires large numbers of subjects, since the groups are independent. Many researchers in other disciplines use the classic experimental design (Groups 1 and 2 only), while others

(IIR included) use a modified design based on Groups 3 and 4. This design is called a *Posttest-only Control Group Design* [47]. Campbell and Stanley [47] argue that these are the only two groups needed, if subjects have been randomly assigned to the groups.

All of these designs rest of the assumption that subjects comprising each of the groups are equal across a range of characteristics. Characteristics which might, if not equally distributed across groups, conspire to generate spurious results — results not caused by the stimulus, but by some other characteristic of the group. Random assignment can be used to increase the likelihood that these characteristics are distributed equally across groups. While it is usually not possible to conduct true random sampling in IIR evaluations, it is possible to randomly assign subjects to groups (or conditions).

6.2 Factorial Designs

Currently, the more common way for researchers to discuss experimental design in IIR is as *factorial designs*. This is particularly useful when studying the impact of more than one stimulus or variable. The models presented above assume a binary stimulus (experimental and control), but the researcher might also be interested in studying the impact of a number of factors[2] on one or more outcome variables. Factorial designs accommodate this. In the preceding example there was one factor, system type, which had two levels, experimental and baseline. If the researcher believed that there might also be differences in the outcome variable based on the sex of subjects, then sex would be an additional factor, with two levels, male and female. This is tightly coupled with the previous levels of measurement discussion; the factors in a factorial design should be discrete. The levels represent distinct categories rather than ratio level values.

There is a specific notation and language for describing factorial designs. If the relationship between the two factors mentioned above were examined (system type and sex) in relation to an outcome measure such as performance, then the experiment is described as a 2×2 factorial

[2] Used as a synonym for independent variable.

2×2 Factorial

		System		
		A	B	Total
Sex	Males			
	Females			
Total				TOTAL

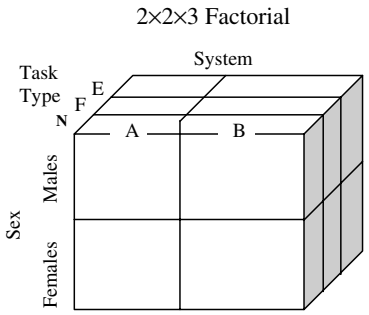

Fig. 6.2 Example factorial designs. The first example is a 2×2 design with one independent variable, system and one quasi-independent variable, sex. The second example is a 2×2×3 design with one additional independent variable, task-type, which has three levels: navigational (n), fact-finding (f), and exploratory (e).

design. Both the number of digits and their magnitudes are meaningful. The number of digits describe the number of factors (system type and sex) and the magnitude of each number describes the number of levels of each factor (experimental, baseline; male, female). If another factor called task-type were added which had three levels (exploratory, fact-finding, and navigational), the experiment would be described as a 2×2×3 factorial. These two designs are illustrated in Figure 6.2. Such illustrations aid in describing a study and allow researchers to understand and communicate the different types of comparisons that are available. The different combinations of levels generate different conditions (in the 2×2 there are four conditions and in the 2×2×3 there are 12 conditions). Each condition will have some value on one or more outcome variables. Comparisons can be made using cell, column and row values. Each factor adds another dimension to the representation and studies with four or more factors do not lend themselves as easily to this type of representation and are not conducted that often anyway because they require large numbers of subjects and it is difficult to interpret results.

6.3 Between- and Within-Subjects Designs

Studies can also be characterized with respect to whether the independent variables are *between* or *within* subjects. This is an important

distinction which should be made in all reports. Between-subjects means that subjects experience only one level of the variable, while within-subjects means that subjects experience all levels of the variable. Studies can be mixed along this characterization: some variables can be between-subjects, while others can be within-subjects. For instance, system type might be a between-subjects variable, while task-type might be a within-subjects variable. This means that each subject would only use one system, but would have to complete all three task-types. Some variables are necessarily between- or within-subjects. For instance, values on the sex variable are completely beyond the control of the experimenter and reside outside of the study. In the classic IIR evaluation study, system type (or interface type) is typically within-subjects to facilitate comparison of the two systems. Otherwise, it is not possible to ask subjects to compare the systems since they would have only used one of them. In other cases, it might be desirable to make a variable between-subjects to avoid exposing subjects to all conditions of the experiment, which might led to contamination.

6.4 Rotation and Counterbalancing

Rotation and counterbalancing are cornerstones of most experiments and evaluations and are often associated with systems and tasks in IIR evaluations [260]. The primary purpose of rotation and counterbalancing is to control for order effects and to increase the chance that results can be attributed to the experimental treatments and conditions. Although *treatment* typically refers to the things that a researcher tests or manipulates (e.g., interface), it also refers to the tasks and topics which subjects execute when engaging in IIR regardless of whether these items are variables of interest. In most IIR evaluations that involve searching, search tasks are necessary in order for subjects to exercise systems. Even though they may not be treated by the researcher as independent variables, they do function as variables and therefore must be controlled. This is typically achieved through rotation.

Two types of designs can be used to systematically rotate variables, the *Latin square design* and the *Graeco-Latin square design* (Graeco is

also spelled Greco). The Latin square design accommodates a single variable, while the Graeco-Latin square design can accommodate multiple variables — it is essentially a combination of two or more Latin squares. To illustrate the different designs, let us assume that we are testing three interfaces using six different search topics and that each user will complete two topics per interface. The task will be held constant and will be a document finding task.

6.4.1 A Basic Design

First, let us look at a basic design with no rotation (Figure 6.3), where rows represent subjects and columns represent interfaces. Topics are represented in the cells of the table. There are two major problems with this design — the first is related to topic order and the second is related to interface order. A Latin square can be used to control for one of these things, but not both. A Graeco-Latin square is needed to control both.

The major experimental confounds that are introduced by this design are caused by order effects. Specifically, learning and fatigue can produce results that are attributable to the experimental design rather than the treatments. As subjects complete each consecutive topic, they learn more about the experimental situation and the experimental system (assuming that the systems are similar). With each topic encountered, subjects potentially transfer what they learn by completing one topic to their interactions with the next topic, which might result in better performance on topics that are presented last, as opposed to first. Since the order of interfaces is fixed, it is also likely that the subject's

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	1, 2	3, 4	5, 6
S3	1, 2	3, 4	5, 6
S4	1, 2	3, 4	5, 6
S5	1, 2	3, 4	5, 6
S6	1, 2	3, 4	5, 6

Fig. 6.3 A basic design with no rotation. Numbers in cells represent different topics.

earlier experiences will impact later experiences. If, on average, subjects perform best with the last interface, it may be a function of a learning effect, rather than the goodness of the last interface. Another problem with a fixed order for topics and interfaces is related to the potential interactions among the topics and the system. Some topics may be easier than others and some systems may do better with some topics than others. If, for example, it was found that Interface 2 was the best, it may be because Topics 3 and 4 were easier than the other topics. In this case, while the researcher may attribute differences to the interfaces, the differences are really caused by the topics.

Fatigue can also impact the results. As subjects engage in more and more searches, they are likely to become fatigued, especially in experiments that last over one hour. At the beginning of the study, subjects may be more motivated and attentive than at the end of the study. When subjects become fatigued they may move quickly through the experiment just to finish. They may become cognitively exhausted and just not be able to perform as well as they did at the start of the study. If, for example, it was found that Interface 1 was the best, it may be because subjects were more energized and worked harder in the beginning of the study than at the end.

6.4.2 A Latin Square Design

To improve the design in Figure 6.3, a *Latin square* can be used to control for the effects of one of the variables — either topic or interface. Latin square designs are used to rotate and control for a *single* variable and in Figure 6.4 this variable is topic. The items in the cells of a Latin square are distinct and should appear an equal number of times in each row and each column. This can be accomplished fairly easily: for each row, topics are shifted among the columns in a systematic way. Since there are six topics, we need six rows (or six subjects) to get through one topic rotation. Note that each user completes all topics. It is also important to note that these designs do not eliminate learning or fatigue, but distribute their impact equally across all treatments and conditions.

While the rotation in Figure 6.4 is balanced with respect to topics, it is problematic for other reasons. There are two important things that

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	2, 3	4, 5	6, 1
S3	3, 4	5, 6	1, 2
S4	4, 5	6, 1	2, 3
S5	5, 6	1, 2	3, 4
S6	6, 1	2, 3	4, 5

Fig. 6.4 Basic design with Latin square rotation of topics.

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	2, 4	1, 6	5, 3
S2	3, 5	2, 1	6, 4
S3	4, 6	3, 2	1, 5
S4	5, 1	4, 3	2, 6
S5	6, 2	5, 4	3, 1
S6	1, 3	6, 5	4, 2

Fig. 6.5 A basic design with Latin square rotation of topics and randomization of columns.

this type of rotation does not address. First, it is possible that there may be some interaction among the topics, such that encountering Topic 4 after Topic 3 makes completing Topic 4 easier. Note that for all rows of the table except one, Topic 4 always follows Topic 3. One can see this visually in the design via the diagonal — this indicates that there is still some order preserved in the table (it is easiest to spot this along the '6' diagonal). One way to address this problem is to randomize the order of the columns (excluding the column headings). One could assign numbers to each of the columns and then use a random number generator to determine the column orders in the rotation. Figure 6.5 illustrates the table once this has been done. The properties of the Latin square are still maintained and topics are no longer completed consecutively. Note that even after randomization of the columns (not topics) it is still the case that each topic will be completed first, second, third, etc. an equal number of times.

The second thing that a standard Latin square design does not address is the order effects introduced by the interfaces. One assumption behind a Latin square rotation is that there is no interaction

between the items represented by the rows and columns. Notice in Figures 6.4 and 6.5 that Interface 1 is always used first, Interface 2 second and Interface 3 third. The previous discussion of order effects as they relate to a fixed topic order also applies to a fixed interface order. Learning and fatigue may conspire to impact the results.

6.4.3 A Graeco-Latin Square Design

The solution to the problem described above is to rotate the order in which subjects experience the interfaces. This can be accomplished with a *Graeco-Latin square* which is a combination of two or more Latin squares. This is essentially equivalent to reproducing the Latin square in Figure 6.4 above three times, each with a different interface order. A single representation of this is displayed in Figure 6.6. In this Figure, the interfaces are now represented within the cells instead of as column headings. The column headings represent points in time (or order) and the rows represent subjects. For instance, the first user would use Interface 1 to complete Topics 1 and 2, and then Interface 2 to complete Topics 3 and 4, etc.

Subjects	Time 1	Time 2	Time 3
S1	I_1: 1, 2	I_2: 3, 4	I_3: 5, 6
S2	I_1: 2, 3	I_2: 4, 5	I_3: 6, 1
S3	I_1: 3, 4	I_2: 5, 6	I_3: 1, 2
S4	I_1: 4, 5	I_2: 6, 1	I_3: 2, 3
S5	I_1: 5, 6	I_2: 1, 2	I_3: 3, 4
S6	I_1: 6, 1	I_2: 2, 3	I_3: 4, 5
S7	I_2: 1, 2	I_3: 3, 4	I_1: 5, 6
S8	I_2: 2, 3	I_3: 4, 5	I_1: 6, 1
S9	I_2: 3, 4	I_3: 5, 6	I_1: 1, 2
S10	I_2: 4, 5	I_3: 6, 1	I_1: 2, 3
S11	I_2: 5, 6	I_3: 1, 2	I_1: 3, 4
S12	I_2: 6, 1	I_3: 2, 3	I_1: 4, 5
S13	I_3: 1, 2	I_1: 3, 4	I_2: 5, 6
S14	I_3: 2, 3	I_1: 4, 5	I_2: 6, 1
S15	I_3: 3, 4	I_1: 5, 6	I_2: 1, 2
S16	I_3: 4, 5	I_1: 6, 1	I_2: 2, 3
S17	I_3: 5, 6	I_1: 1, 2	I_2: 3, 4
S18	I_3: 6, 1	I_1: 2, 3	I_2: 4, 5

Fig. 6.6 A basic design with Graeco-Latin square rotation for topic and interface.

Note that this design has the same problem as the design in Figure 6.4: Interface 2 always follows Interface 1 except when Interface 2 is first. To address this problem, the same column randomization strategy described above can be applied. The column randomization should be applied after the Graeco-Latin square has been built, otherwise it cannot be ensured that each topic will be paired an equal number of times with each system.

Randomization should be used to assign subjects to the different rows in the table, even when the columns have been randomized. All experimental designs assume random assignment of subjects to conditions. To accomplish random assignment, numbers could be assigned to the rows in Figure 6.6 and a random number generator could be used to determine the order of the rows. Random assignment to condition controls for any potential differences that might be attributable to subjects. The assumption is that any individual differences in subjects (e.g., intelligence, search experience, and motivation) that might impact the results will be equally distributed across condition and therefore controlled as much as possible.

Notice that the rotation in Figure 6.6 provides insight into how many subjects are needed for the study. We know that we need at least 18 subjects to get through the rotation once and to keep the study completely balanced we would need to recruit subjects in batches of 18. However, this is not the only way to determine an appropriate sample size. Statistical power, representativeness and generalizability are also important factors.

6.4.4 Using the Mathematical Factorial to Construct a Design

Another method that can be used to construct an experimental design makes use of the mathematical factorial to enumerate all possible orders for topics and interfaces. However, it is important to note that this is *not* a Latin square rotation — it is a factorial rotation. It is also important to note that this type of rotation is infeasible and cannot be used to create a completely balanced designed in most cases. For instance, in our example with three interfaces and six topics, we would first need

to do a factorial for interface type ($3! = 3 * 2 * 1$), which results in six possible orders. Next, we would need to do this for the six topics ($6! = 6 * 5 * 4 * 3 * 2 * 1$), which results in 720 possible orders. To make the experiment completely balanced, we would need 4320 subjects ($6 * 720$). It is unlikely that anyone would have the resources to recruit and study 4,320 subjects for a single study and besides, studying this many subjects is not really necessary since at some point statistical power plateaus. One might select a portion of these orders, but this would not result in a completely balanced design. However, there are some types of situations, where a factorial rotation is feasible. For instance, two interfaces ($2! = 2 * 1$) and four topics ($4! = 4 * 3 * 2 * 1$) results in 48 possible orders.

6.5 Randomization and User Choice

Another method that can be used to create experimental rotations is to randomly create orders. This can be done by combining different orders of the interfaces and topics, or in conjunction with a Latin Square, where the main variable of interest, interface type, is rotated using a Latin Square and topics are randomly assigned to subjects. It is often the case that researchers want to include more topics in a study to increase generalizability and randomization is selected as a way to assign topics to subjects. However, topics are unlikely to be equally represented in the data set (unless very large numbers of subjects are studied). Thus, results may be attributable to topics and/or topic interactions with other independent variables. If one can use a Latin or Graeco-Latin Square design, then it is a better choice for ensuring a more balanced experimental designed.

Another approach is to give subjects a choice of topics. For instance, subjects might be presented with 10 topics and allowed to select four that they would like to research using the experimental systems. The justification for this approach is it helps increase subjects' motivation [292]. However, if one does this, one should be careful not to give subjects too many choices and have some control over how many topics are completed with particular systems. The danger in letting people choose is that the choices may naturally create a situation where topic

effects are present. There will likely be an unequal distribution in subjects' choices, resulting in some topics being overrepresented in a study and others being underrepresented. It may also be the case that some system–topic pairs occur more frequently unless extra effort is taken to prevent this.

6.6 Study Mode

The mode in which a study is administered can also vary. IIR evaluations can be administered in batch-mode, where multiple subjects complete the study at the same location and time or in single-mode, where subjects complete the study alone, with only the researcher present. The choice of mode is ultimately determined by the purpose of the study. In studies where subjects are deceived in different ways, completing different sequences of activities or will be interviewed, single-mode studies are more appropriate. If the experiment is relatively self-contained, subjects do similar things and can be directed via computer, then batch-mode is appropriate.

Batch-mode studies are very efficient — more subjects can be ran in a shorter period of time. However, it is important to note that subjects can influence one another even when they do not communicate verbally. For instance, in a batch-mode design, the first person to finish the study will likely signal to others that the end of the experiment is approaching. As a result, the remaining subjects might work faster and be less thoughtful, even if they are in different conditions that require more time. Thus, one should think carefully about non-verbal signals that are present in batch-mode studies, what these signals might communicate and how they might contaminate or change a subject's experiences and subsequent behaviors.

Studies can also be administered via the Web instead of in the laboratory. Toms et al. [272] adapted the traditional TREC Interactive Track IIR evaluation model so that it could be run on the Web. The WiIRE framework provided an infrastructure where researchers could plug-in different systems or interfaces for evaluation and tailor common instruments, such as questionnaires, to their needs. Researchers are increasingly experimenting with different ways of administering

evaluations online, although the impact of this on the quality of the evaluation data is unclear. The main concern is that allowing people to login to a system and complete a study in any environment potentially introduces confounding variables that will be unknown to the researcher. For instance, one subject might complete the study while sitting in a loud environment, another might multi-task between the study and other tasks, including text messaging and emailing, while another might solicit help from others or refer to alternative resources of information while searching. Of course, these all represent real use scenarios (and subjects can be instructed about what is expected of them) and this may be what interests the researcher. However, such studies should not be treated as controlled experiments, because they are not.

6.7 Protocols

A study protocol is a step-by-step account of what will happen in a study. It is useful to have a document describing in detail exactly what should happen to guide the researcher. Check lists and other such documents can be used to ensure consistency in the administration of the study. This consistency helps maintain the integrity of the study and ensure that subjects experience the study in similar ways. Creating a detailed protocol also helps ensure that the experiment will run smoothly and that the researcher knows what to expect. In cases where multiple researchers are conducting a study, a protocol helps ensure that the same steps are followed for each subject.

6.8 Tutorials

When subjects encounter new IIR systems it is often the case that they need some instruction on how to use them. Many of the systems that IIR researchers investigate are experimental and thus, differ from the standard systems to which subjects are accustomed. In the past, researchers have created print tutorials to introduce subjects to an experimental system, while others have verbally administered tutorials. Of the two, the print option is best because it ensures that the

presentation is consistent. These days, an easy method for creating a tutorial is to record a video tutorial using screen capture software. This video can be played for each user and it can be guaranteed that what is told and how it is told is consistent. It is best to first develop a script before creating a video tutorial.

There are objections to the use of tutorials and other instructional materials on the grounds that they potentially bias subjects and that in real life people do not read instruction manuals. The issue related to bias is arguably the more important objection; the tutorial may suggest to subjects how they should interact and behave. If one is using a measure such as uptake of a new feature and the feature is prominently discussed in the tutorial, the measure may just reflect how cooperative subjects are, rather than their real interests in the feature. However, if the purpose of the experiment is to evaluate a new feature, then asking people to use the feature seems reasonable since it must be used in order to be evaluated. When it is necessary to provide a tutorial researchers should ensure consistency and balance in the presentation and consider how this experience might influence subjects' behaviors and the study results.

6.9 Timing and Fatigue

Another issue that needs to be considered is the length of time the study will last. This is a critical issue because typically subjects are performing activities that take some length of time to complete. Unlike studies in psychology, where hundreds of trials can be conducted in a single hour, very often only four search tasks can be completed in one hour. Moreover, search activities can be exhausting both mentally and physically. There are no set rules on how long one should give subjects to complete tasks; this is usually contingent on the type of task and study purpose. For instance, in an evaluation of Web search result surrogates, Käki and Aula [158]) imposed short time limits in an attempt to simulate how people actually scan surrogates in real life. In many other evaluations, subjects are given 10–15 min to complete search tasks.

6.10 Pilot Testing

One way to get an estimate of how long a study will last is to conduct a pilot test. Pilot tests help researchers do a number of other things besides estimate time. They help researchers identify problems with instruments, instructions, and protocols; allow systems to be exercised in the same way they will be in the actual study; provide researchers with an opportunity to get detailed feedback from test subjects about the method; help researchers gain comfort with administering the study; and finally, they can be used to train inexperienced researchers. Ultimately, pilot tests help researchers identify and eliminate potential confounds and errors that might otherwise compromise the integrity of the study results.

7

Sampling

Many different items are sampled in IIR evaluations including users, tasks, topics and documents, although the biggest emphasis is on users. The term *element* will be used to refer to items that are sampled, although in most of the discussion that follows users are the focus.

First, it is important to note that it is generally not possible to include all elements from a population in a study, which is one of the main reasons for sampling. In most cases we do not know or have access to all elements in a population. Thus, populations are sometimes described as *theoretical*. It is to this theoretical collection of elements that researchers aim to generalize their results.

The population is usually not mentioned in most IIR reports. Instead, it must be assumed that the target population is all people who engage in online information search, or all literate people who engage in information search between the ages of 18–70 or just all people. There is an implied population behind all studies that involve samples (whether the elements are people or tasks or documents), even if it is not stated explicitly.

Any discussion of sampling must include a discussion of the specific numbers of elements that should be included. In other words, what

is a sufficient sample size? There is no principled, all-encompassing response to this question. There are some general rules of thumb that can be used to estimate how many subjects one needs based on one's design. For instance, counter-balancing a design will show how many subjects are needed to get through one rotation. There are also formulas that can be used to conduct power analysis in cases where a factorial design is used to determine how many subjects are needed in order to achieve a specific power[1] and formulas for survey research to determine appropriate sample sizes given specific margins of error and confidence levels.[2] In general, more is better, although there is a point at which one receives diminishing returns with respect to statistical power. Sample size is used in almost all statistical computations and if the sample is small, then it will be difficult to find statistically reliable results, unless the effect is extremely strong.

It is important to note that some types of methods, in particular qualitative methods, do not rest on the notion of probability sampling and are not focused on statistical testing. It is often the case that qualitative studies and some naturalistic studies have small numbers of subjects. There is a trade-off between the number of subjects that can be included and the intensiveness and depth of the interactions that can occur with those subjects.

Ultimately, the purpose of the research determines the sampling approach and the sample size. The important thing is for researchers to understand the limitations associated with their sampling strategies and exercise caution with interpreting results and generalizing findings. One assumption behind inferential statistical testing is that the sample is representative of the population from which it was drawn and that it was drawn using probability sampling techniques (although this is a theoretical assumption because in most cases this is violated). One should be mindful that a statistic is an estimate of some value in the population (known as a *parameter*); if this estimate is made on the basis of an unrepresentative sample, then statistical test results will not be reliable.

---

[1] See http://www.stat.uiowa.edu/~rlenth/Power/ for an example.
[2] See http://www.surveysystem.com/sscalc.htm for an example.

There are two major approaches to sampling: probability sampling and non-probability sampling. With few exceptions, most sampling in IIR is of the non-probability variety because it is nearly impossible to meet all of the criteria of probability sampling, especially when sampling people.

7.1 Probability Sampling

Probability sampling suggests ways of selecting a sample from a population that maintains the same variation and diversity that exists within the population. If there were no differences among people and all people were exactly alike, then there would not be a need for sampling; in fact, a researcher would only need a sample size of 1. However, since people vary along a number of characteristics (e.g., sex, level of education, and search experience), the goal of probability sampling is to create a sample that contains the same variation of these characteristics that exists within the population. Such a sample is termed *representative* because the distribution of these characteristics in the sample matches their distribution in the population. For example, if females comprised approximately 60% of the population and males 40%, then a representative sample would contain roughly 60% females and 40% males. The important thing about representative samples is that they increase the *generalizability* of the results. Generalizability is related to the extent to which the study results reflect what would happen in the entire population.

Although it is nearly impossible to have a perfectly representative sample, a probability sample is more representative than a non-probability sample. Probability sampling rests on the assumption that all elements in the population have an equal chance of being selected. There are several techniques that can be used to accomplish this, most notably simple random sampling. However, all of these techniques rest on the assumption that all elements in the population are known and that all elements will be included when selected. Stated another way, in order for each element to have an equal chance of being selected for the sample, it must be known *a priori*, otherwise it does not have a chance of being selected. If one's population is all students at a particular

university, then this information would likely be available. However, if one's population is all adults between the ages of 18 and 60, then it unlikely that a comprehensive list of such elements exists. In most sampling situations, especially those in IIR evaluations, knowledge of all of the elements is not the norm, especially when people are being sampled. This is also the case if one considers the world of possible tasks, topics and documents that exist — it would be impossible to enumerate all of these in order to execute probability sampling. However, if a small corpus of documents functioned as the population (e.g., a TREC corpus), then it would be possible to use probability sampling to randomly select documents.

The other problematic piece of this is that all selected elements would need to be included in the sample. With documents or other inanimate objects, this is not problematic so long as the researcher has access to these items. With human beings, this is more problematic since people cannot be forced to participate in a study. People who decline to participate might possess some characteristic that distinguishes them from those who agree to participate. This characteristic will then be absent from the sample which will in turn compromise the representativeness of the sample to randomly select documents.

As mentioned earlier, in order to conduct probability sampling, every element must have an equal chance of being selected. This assumes a constant probability and that one either draws the total sample simultaneously or that sampling is done with replacements. Otherwise, those elements selected later will have a greater chance of being selected than those selected earlier, since the probability of being selected changes as each element is removed from the population and added to the sample.

The other possible method for ensuring that all elements have an equal chance for inclusion is sampling with replacements, although this is really only a theoretical solution. Sampling with replacements means that each element that is selected is returned to the population and essentially exists in two places, the sample and the population. For instance, if one were drawing names from a hat, this means that after a name is drawn, one would note that this element was part of the sample and then return it to the population (i.e., the hat) and continue

drawing more elements. This ensures that each element always has the same chance of being selected, but this also implies that some elements can be selected more than once. Of course, it is not practical to include the same person in the same study more than once, so sampling with replacements cannot be used when each element in the sample needs to be unique.

Another factor affecting the representativeness of the sample is its size. Even when one is able to employ probability sampling, if enough elements are not selected, then the sample will not be representative of the population. For instance, if a population size is 1000 and a researcher randomly selects 10 elements for the sample, then it is unlikely that the variation that exists in the population will exist in the sample. In this example, the sampling ratio is 1%. However, the relationship between a sample and population size is not linear with respect to statistical power. At some point, statistical power begins to plateau and each increment in the sample size adds virtually no statistical power.

There are three major probability sampling techniques: simple random sampling, systematic sampling, and stratified sampling. *Simple random sampling* is the most basic type of probability sampling. First, one creates a list containing all elements that are to be sampled and associates numbers with each element. For instance, if items are listed in a spreadsheet, the row numbers could function as numeric identifiers. Next, one uses a random number generator or a random number table to identify which elements will be included in the sample. It is assumed that most readers are familiar with these techniques. The second technique, *systematic sampling*, is also likely to be familiar to readers. When using this technique, every kth element in the list is selected for the sample. K can be determined by dividing the population size by the desired sample size.

Stratified sampling is a technique that can be used in conjunction with simple random sampling or systematic sampling. The purpose of stratified sampling is to subdivide the population into more refined groups, which are defined according to specific strata, and then select a sample that is proportionate to the population with respect to the strata. For instance, if one were sampling documents that were

associated with specific topics such as science, technology, and literature, one might want to ensure that the proportion of documents associated with each topic in the sample was equal to the proportion of documents associated with each topic in the population. In this example, there would be one strata, topic. After a desired sample size has been determined, proportions can be used to determine target numbers of documents for each topic. Elements from each group would then be sampled using a random number generator, with the number of elements selected for each sample group proportionate to the size of these groups in the population. It is possible to include multiple strata (e.g., topic, publication source, date). This would require the researcher to identify target proportions at a finer level. Multi-dimensional matrices are helpful for visualizing multiple strata where each cell will have a specific proportion associated with it.

Systematic sampling can also be used in conjunction with stratified sampling. Elements of the population can be grouped and sorted according to the various strata in a list format, rather than a matrix. For instance, documents might first be sorted according to topic, then publication source and then date. Once the list has been assembled and sorted accordingly, systematic sampling can be used to select elements from the list. Specific proportions do not have to be associated with the strata because an equal proportion of elements should be selected since the list is ordered.

7.2 Non-Probability Sampling Techniques

In most IIR evaluations, non-probability sampling techniques are used. There are several reasons for this. Researchers often do not know all of the elements in a population and therefore cannot generate the lists required for probability sampling. Even if the elements are known, researchers may not have access to them. For instance, some people may refuse to participate or some documents may be impossible to obtain. Financial constraints and other resources also limit what is possible. Even if one were able to select a random sample of people in a geographic area for an IIR study, it is unlikely that the project budget

would be large enough to pay the travel and lodging costs for potential subjects.

The biggest weakness of non-probability sampling techniques is that they limit one's ability to generalize. This does not mean that research using non-probability sampling should be dismissed, but it does mean that researchers should exercise caution when generalizing from their data and be explicit about sampling limitations in their reports.

There are three major types of non-probability sampling: convenience, purposive, and quota. *Convenience sampling* is the most common type of sampling used by researchers in a number of areas, including IIR. This type of sampling is used in IIR and IR more broadly, to sample users, topics and documents. Convenience sampling is relying on available elements to which one has access. When researchers recruit undergraduate students from their universities or people that are geographically close to them, this is convenience sampling. When newspaper articles or congressional reports are used to create a corpus because one has access to these documents (including copyright permissions), this is convenience sampling. These represent smaller subsets of what is possible, but this is usually a result of practical constraints not researcher negligence. Even the massive log-based studies conducted at search engines companies rely to a certain extent on convenience sampling, since they only include users of a particular search engine. Of course, they have a lot of users, so this helps, but it still is not a probability sample since people choose to use one search engine over another.

Purposive or *judgmental sampling* happens when a researcher is interested in selecting subjects or other elements that have particular characteristics, expertise or perspectives. Inclusion and exclusion criteria are usually associated with purposive sampling; such criteria indicate who may be included in a study. For instance, during an initial evaluation of a new IIR interface, one may purposively select usability experts or students enrolled in a human–computer interaction course to gain very detailed and critical feedback about the interface. Subjects without such expertise may not be able to consider and articulate as wide a range of responses as those with such expertise.

The last type of non-probability sampling is *quota sampling*. It is identical to stratified sampling except that the technique for populating

the strata (or cells) is a non-probable technique (most likely convenience sampling). For instance, after one has created the strata and identified the target number of subjects for each stratum, during recruitment, cells of the table are populated on a first-come-first-served basis. The researcher might send a solicitation to all undergraduates and ask them to characterize themselves according to these strata when replying to the solicitation. The researcher would then be able to ensure that the appropriate numbers of subjects with each characteristic and combination of characteristics are included in the sample. The other major difference between quota sampling and stratified sampling is that the proportions associated with each stratum are usually not as accurate for quota sampling as they are for stratified sampling since information about such proportions in the popular may not be available.

7.3 Subject Recruitment

There are many different methods for recruiting subjects including posting signs, sending solicitations to mailing list, inviting those who work in one's organization, using subject pools and using referral services. Researchers also work in conjunction with others to identify and recruit subjects when they do not have immediate access to target subjects; for instance, a researcher might work in conjunction with the military to recruit intelligence analysts. Recently, many researchers have relied on crowdsourcing and mechanical Turk to recruit people to make relevance assessments via the Web [7]. As more people begin to experiment with conducting IIR evaluations on the Web, additional recruitment strategies may be observed such as Web advertising, mass mailings and virtual postings in online locations.

Because many researchers rely on convenience sampling and are located in academic institutions or technical–industrial settings, there has always been an overrepresentation of undergraduates, computer science students and researchers, and library and information science students in published studies. As long as one describes faithfully the sample and the recruitment techniques, then readers can make their own determination about the validity and generalizability of the results.

Having such groups as subjects is not inherently bad, but extrapolating wildly from results is. The more important question is about the extent to which these groups may be biased. In particular, many researchers have used their lab mates or those in their research group or even themselves as study subjects. This is very problematic because these people likely know something about the purpose of the study and the desired outcome. When study subjects also end up analyzing the data, there is even more room for concern.

7.4 Users, Subjects, Participants and Assessors

There are a variety of names given to humans who are studied in IIR research. The most common are users, subjects, research participants and assessors. A general rule of thumb for distinguishing among these different labels is as follows. The term *user* is often used in situations where those being studied are actual users of a system. For instance, in log-based Web studies the data are usually generated by real users. These users are not using the system for the sole purpose of generating research data (indeed, most probably do not even realize how their data are used), but instead the data is a byproduct of their normal use of a system.

Interestingly, the phrase *user study* was originally used in information science to describe studies that investigated people's information seeking needs. Siatri [242] traced the first user studies in information science to the 1948 works by Urquhart [278] and Bernal [26] who studied the distribution and use of scientific information, and the reading habits and needs of scientists, respectively. These days user study is used more generally to describe any study that involves human participants, which really dilutes its meaning.

The terms *subjects* and *research participants* are used to describe people who knowingly participate in a research study. Subjects and research participants are the subset of the user population that has been selected for inclusion in a study. The sole reason that these people are using a system is because they are part of a study. The term subject has typically been associated with laboratory studies, while research participant has been associated with naturalistic and qualitative studies.

It is also the case that some people dislike the term subject as it is said to be dehumanizing.

The term *assessor* has been used to describe people whose sole purpose is to make relevance assessments. There is a fine line between an assessor and a subject; one could justify the distinction by noting that the only data produced by assessors that are of interest are the relevance assessments. Although traditionally the behavior of assessors has been of little concern, several researchers have started to investigate assessors [14, 226], thus treating them more as research subjects.

8

Collections

8.1 Documents, Topics, and Tasks

Most IIR evaluations require subjects to search for information. Thus, one important consideration is the identification of a set of documents (or more generally, information objects) for subjects to search and a set of tasks or topics which directs this searching. Along with these things, a researcher must also make some decisions about how the relevance of the information objects to the topics will be determined. These items — corpus, topics, and relevance judgments — comprise what is commonly known as a *test collection* in IR. In IIR evaluations, these items can be thought of as instruments just like questionnaires and logging software.

Although *collection* is often used as a synonym for *corpus*, in this paper these words will be used to indicate two separate things in the way that they are used in the context of TREC [121]. A *collection* consists of topics, a corpus, and relevance judgments. A *corpus* is the set of documents, or information objects, that subjects access during a study. IIR evaluations vary on the extent to which they have each of these components. In some studies, subjects are provided with standard

topics or tasks, search the entire Web and make the final relevance judgments. In other studies, subjects are provided with topics, asked to search a specific corpus and although they are able to make relevance judgments, these are ultimately compared against a gold standard (e.g., TREC relevance assessments).

8.1.1 TREC Collections

TREC collections, and specifically those used by the Interactive and HARD Tracks, have been used in a number of IIR evaluations. If one is conducting a controlled, laboratory study or system evaluation, these collections are attractive for a number of reasons. There is a finite and theoretically knowable set of documents and some information is available about the number of documents that are relevant to different topics. It is also the case that relevance assessments exist and different kinds of performance metrics can be computed.

There are also a number of limitations associated with using TREC collections in IIR evaluations. It is generally known that users' queries retrieve different documents than the batch queries used in system-centered evaluations, so it is possible that subjects will find documents that were not included in the relevance pools [129]. If a document was not in the pool, then it would not have been judged by the original assessor. Strictly speaking, documents in this situation are considered not relevant. One might be tempted to just independently assess these documents, but this potentially perturbs any findings since the majority of assessments will be from a single assessor and a small number will be from people who did not experience the original assessment context. The stability of the TREC collection rests on the assumption that the relevant documents are relevant to a single user at a single point in time. Mixing assessments made by others violates this assumption.

A bigger problem is related to the extent to which relevance assessments generalize. Numerous studies have demonstrated that relevance assessments do not generalize across subjects [80, 129]. Indeed, it is understood that different people will make different relevance assessments given the same topics and documents. TREC acknowledges

this and makes no claims about the generalizability of the relevance assessments [288]. Their stance is that these assessments represent one user's judgments at one point in time. This is fine when one is comparing relative system performance, but problematic when trying to use these assessments as benchmarks in studies with new subjects. It has even been argued that the performance measures that are computed in IIR evaluations using TREC relevance judgments actually represent the extent to which subjects agree with the assessors, rather than performance [80].

An alternative is to completely ignore the relevance assessments made by assessors and instead generate new relevance assessments based on subjects' actions. This will impact one's abilities to compute certain performance metrics, especially those based on recall, but these may not represent performance concepts that are well-suited to the purpose of the study anyway. Creating new assessments can be done using a consensus approach [301] or by just accepting what individual subjects identify as relevant for the topic. One problem with consensus-based approaches is that one will likely end up with a lot of documents that were saved by a single user, but discarded by others. However, in cases where researchers are using graded relevance, a consensus-based approach provides a useful way to grade documents.

Another limitation of using TREC collections in IIR evaluations is that most of the corpora are newswire text. Although it is the case that a large number of internet users get news online[1] this is not the only type of searching task that is performed and more often than not, users browse the same news sources daily rather than actually search for news articles. More recent TREC Tracks have explored different kinds of corpora, such as blog and legal.

Finally, another major criticism of using TREC collections in IIR evaluations is that the topics are artificial. However, it is possible to make this criticism of any study that uses artificial topics and tasks, so this is not unique to TREC collections.

[1] According to the latest report from the Pew Internet & American Life Project Tracking surveys 72% of online users view news online http://www.pewinternet.org/trends/Internet_Activities_6.15.07.html.

8.1.2 Web Corpora

In many studies, the Web is used instead of a corpus. Subjects are allowed to search the Web and there are no constraints on what pages and resources they can use. The major drawback to using the Web is that it is impossible to replicate the study since the Web is constantly changing. Two subjects in the same study might issue the same query at two different points in time and get completely different results. Although network delays are less of an issue these days, it is also the case that different subjects might experience different reaction times depending on the time day. Factors such as authority and quality enter more into subjects' relevance assessments since these are not controlled for as they would be in a closed collection. If the subject knows that all documents are from *The New York Times*, then the subject does not have to worry as much about source credibility. This suggests that subjects should be required to provide more and different kinds of relevance assessments when searching the web, than when searching closed corpora [271]. It is also the case that subjects may base their relevance judgments on more things than just the text (e.g., design, style, and images). Finally, with open web searching, researchers are only able to use a small number of the well-established performance metrics.

In log-based Web studies, corpora are often built on-the-fly, as subjects visit the Web pages. For researchers working at search engine companies, these corpora have traditionally consisted of the search engine home page, the search results page and first level search result. Recent browser plug-ins allow these researchers to gather page visits that go beyond this traditional tripartite set, but some newer plug-ins have the potential to corrupt this data.[2] Regardless of who is conducting the study there is potential to create an offline corpus of documents viewed by study subjects and to perform future experiments with this collection (cf. [269]).

It is also possible to study IIR in the context of open or closed corpora that are accessible via the Web. For instance, one might study IIR in the context of a digital library, a proprietary database, an internal business corpus or intranet. Searching such circumscribed corpora

[2] http://mrl.nyu.edu/∼dhowe/trackmenot/

provide more control over the retrieval situation and a greater understanding of the corpus and range of document types. Assumptions can also be made about the quality and kind of documents within the corpora. However, the difficulty is ensuring that the corpus is of a sufficient size so that the retrieval task and results are meaningful.

Finally, there are some closed Web corpora that have been crawled and assembled by various research groups, including the TREC .gov corpus. These corpora share some of the positive things related to closed corpora, mainly they are stable, theoretically definable and self-contained, and allow for the computation of recall-based measures.

8.1.3 Natural Corpora

Natural corpora are corpora that have been assembled over time by study participants. Studies using natural corpora are most often in the context of personal information management [119]. The benefits of using such corpora are that they add to the realism of the study and allow subjects to interact with documents that are meaningful to them and with which they are familiar.

The use of natural corpora also has drawbacks. The biggest problems are lack of replicability and equivalence across participants. Each subject's corpus will differ in size and kind. One subject may have a small set of word processing documents and Web pages, while another subject may have gigabytes of documents representing a variety of file types. Thus, the researcher needs to know something about the number and kind of files available on the machine to interpret the subject's behavior. Cross-user comparisons are difficult since each subject's experience will be different and dictated in part by their own corpus. There may also be unknown document type–tool interactions. For example, the experimental tool may handle some document types better than others and there is no way to control the distributions of such document types across subjects' corpora. Furthermore, if the corpus resides on the subject's machine, the subject may prepare for the study by organizing, deleting and filing things, which changes the natural state of the corpus. Finally, researchers are unable to do follow-up experiments since the corpus resides on the subject's machine and changes constantly.

To overcome some of these problems, natural corpora can be transferred to a research machine where they can be controlled. This might be necessary if the IIR application being tested is not robust enough to be deployed in other environments. The downside to this is that subjects will have to spend time preparing and transferring these corpora to the researcher. Of course, there is also the possibility that some organizing and self-censoring will occur because subjects may be self-conscious about how their corpora exist in natural form. As users accumulate more and more electronic information and increased attention is given to personal information management, we will likely see more researchers developing methods for using and experimenting with natural corpora.

8.1.4 Corpora and Behavior

One important thing to point out about choice of corpora is that it will impact the way in which subjects behave and seek information. For instance, behavior with a Web-based corpus with lots of hypertext links will be different from behavior with a TREC newspaper corpus that has no hyperlinks. The search strategies and tactics that subjects employ when using their own personal corpora might differ from those they employ when searching the open Web. When designing an IIR study, it is important to recognize the potential impact of the corpus on behavior and interactions.

8.2 Information Needs: Tasks and Topics

A user's information need is perhaps one of the most critical aspects of information seeking and retrieval. This need forms the basis of the user's activities and relevance judgments. Much has been written about the nature of information needs and it is generally accepted that people often have a difficult time articulating their information needs and translating them into a vocabulary that is appropriate for a system [19, 20, 25, 104, 267, 297]. Research has also shown that information needs evolve during the search process; this evolution results in dynamic relevance assessments — that is, as people learn more about their information needs, their relevance behaviors change [266, 281].

Although it is difficult to create an all-encompassing definition of an information need, most information needs can be characterized in terms of task and topic. In the published literature, these three terms (information need, task, and topic) are often used interchangeably. However, it is important to distinguish among them so that one is clear about what is being studied. A task represents the goal or purpose of the search — this is what a user wants to accomplish by searching. For example, a traditional task is gathering information to write a research report. Other tasks include planning travel, monitoring sports scores, navigating to a homepage, or re-finding previously seen information. The topic represents the subject area that is the focus of the task. For example, one might be gathering information to write a research report about the malaria epidemic in Africa or one might be planning travel to Australia. One task might be associated with several topics and one topic might be the focus of many different tasks. It is the combination of the specific task and topic that forms the information need.

Historically, IR focused on the topical aspects of the information need. In the early years, systems were developed for trained searchers who were searching on behalf of a client, typically a researcher or scientist looking for exhaustive information about a particular topic. The task was often constant (or assumed to be constant) and was more recall-oriented. Even after the target user group changed to include non-expert searchers, the model search task was still somewhat stable since IR systems were only located within specific environments. With the development of Web IR, different kinds of task models began to develop as the types of users, tasks and use environments diversified. Web IR, in particular, has brought to the forefront precision-oriented information needs, where users are looking for one or a small number of documents rather than all of the documents about a particular topic. Currently, there is a swing back towards recall-oriented information needs, including exploratory tasks, where users are looking for a larger number of documents and have information needs that are unfocused and evolving [190].

Although research trends can impact which tasks are considered important, underlying all IR research is some *user model*, which is an abstract representation of target users, and one or more *task models,*

which represent the goals of users. One user model that has been used a lot in IR is that of a librarian or other search intermediary. Other examples include intelligence analysts and undergraduate students. Examples of task models include finding documents for a survey article, homepage navigation, and fact-checking. The user and task models help define the particular behaviors and activities the IR system is intended to support and help determine the appropriateness of particular evaluation measures and study participants. Although studies may have one or more task models, typically there is only one user model. User and task models are often implied and inherited from research tradition. For instance, underlying many IIR evaluations is the TREC *ad-hoc* task which is modeled after exhaustive searching.[3] Appropriate user models include students writing a survey paper for a class, or intelligence analysts preparing a briefing for a military official. Historically, user and task models have been held constant within a particular study and the only thing that changed was the topic (hence the reason for referring to *TREC topics*), but current research is starting to employ different task models within a single study and using task as an independent variable. Some common task-types that have been investigated include navigational, known-item, fact-finding, resource finding, homepage finding, exploratory and informational.

Task has been shown to affect users' information seeking behaviors and relevance judgments in a variety of ways and is a good candidate variable for understanding more about search systems and user behavior [163, 170, 171, 249, 266, 281]. Vakkari [279] provides an overview of task-based information searching. Byström [43, 44, 45] has conducted a large number of studies investigating task complexity and how tasks can be defined, measured and studied. In particular, Byström and Hansen [44] distinguish among work tasks, information seeking tasks, and information retrieval tasks. Li and Belkin [177] developed a faceted approach to the conceptualization of task. Ingwersen and Järvelin [139] have also provided task classifications. Bell and Ruthven [24] and Gwizdka [114] explore measures of task complexity and difficulty. Toms et al. [273] examined the effects of several task-related variables on interactive

[3] All TREC collections (and Track) have some user and task model associated with them.

search behavior. Kim and Soergel [172] identified various task characteristics and proposed how they can be used as independent variables in research studies.

Search behavior and relevance judgments can also vary according to topic, but usually this is a result of variations in user related variables, such as how much a user knows about a particular topic (e.g., [134, 291]), and corpus related variables, such as how many relevant documents are available about a particular topic (e.g., [135]). It is common in many IR and IIR evaluations to investigate performance and other dependent measures with respect to topic in a *post-hoc* fashion (i.e., at the end of the study), but it is uncommon to treat topic as an independent variable.

8.2.1 Generating Information Needs

Creating information needs for a study is difficult for a number of reasons. It is not always clear at what level of specificity a task or topic should be defined or how many facets should be used to describe the need. For instance, tasks such as gathering information to write a research report or planning travel can be broken down into a series of sub-tasks or, conversely, they can be grouped together into the broad task of seeking information. A topic such as elephant poaching can be further broken down into techniques, places, penalties, policies, etc. or it could be described at a higher level simply as elephants. Other considerations that must be made is whether there is a sufficient number of documents in the corpus and whether target users will have the basic abilities and knowledge to complete specific tasks. One of the most difficult aspects of creating information needs is ensuring the needs are appropriate to what is being studied, but are not over-engineered to guarantee success. Unfortunately, there is little formal guidance for creating tasks and information needs and it can be argued that it is impossible to create artificial information needs, since information needs are generally believed to reside within a person's head. Indeed, most research focuses on information tasks, rather than information needs. Elsweiler and Ruthven [86] outline some steps that one might execute to create search tasks in the domain of PIM.

In many IIR evaluations, where information needs are assigned to subjects, researchers are using TREC collections, which come with topic descriptions. This is one benefit of using TREC collections. However, there are times when researchers must create information needs, especially if they are studying the open Web. A common approach is to examine query logs and work backwards from the queries to develop information needs. It is important to note that queries are not synonymous with information needs. Often users will issue a number of queries during the resolution of an information need. This has been one criticism of log-based studies that only passively monitor what users do — such studies amass large amounts of queries, but it is unclear to what end these queries were written. The ability to isolate and study search sessions which might be comprised of a number of queries from a single user helps address this, but it still does not allow one to understand the nature of the information need. Despite the difficulties of going from a set of queries to an information need, such queries provide researchers with some insight into the kinds of things for which people are searching.

Another way to develop information needs is to work with experts who would either be assigning such information needs to target users as work tasks, or have the same kinds of information needs [138]. For example, if a system were designed to support intelligence analysts, then a group of intelligence analysts might be enlisted to help develop the search tasks. A final way to approach information need (or task) creation is to first identify different characteristics of an information need — for instance, information needs can be well-defined or ill-defined, fleeting or persistent — and then combine the facets in different ways to construct the needs (e.g., [177]).

8.2.2 Simulated Work Tasks

One important development in IIR evaluation and experimentation has been the simulated work task, which Borlund [34] describes as a short cover story that describes the situation leading to the information need. Simulated work tasks go beyond simple topic-based descriptions of needs by providing more contextual information that is tailored

towards target users. Borlund and Ingwersen [35] note that the simulated work task describes the following to the user: the source of the information need, the environment of the situation, and the problem which has to be solved. This problem serves to make the test person understand the objective of the search. Such descriptions are further proposed to provide a basis against which situational relevance can be judged.

Simulated work tasks are comprised of two major parts, the simulated work task situation and the indicative request (although the indicative request is identified as optional). Together they are called the simulated situation. An example is shown below in Figure 8.1. Notice in this description, the task situation is tailored to the target users — university students. Also notice the indicative request, which provides an example of what subjects might search.

One of the primary rationales for developing simulated work task situations is the criticism that assigned search tasks are artificial, that subjects may not have a context for executing the task and making relevance judgments and that subjects may simply be unmotivated to search for artificial tasks. Of course, the reason for assigning tasks to subjects is to control the search situation and produce conditions that allow for comparison. Borlund's work provides an empirically validated way to use assigned tasks while also personalizing them. This simulated work task method calls for members of the target user group to generate the tasks to ensure that they are relevant to study subjects. This aspect is often overlooked by those using this method to create tasks for their own studies.

Simulated Situation

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

Fig. 8.1 Example of a simulated situation/simulated work task situation [32, 33].

Other researchers have tried to address the motivation problem by presenting subjects with a choice of assigned tasks [153, 292]. When using TREC topics, motivation and requisite background knowledge for making relevance assessments become crucial issues since many of these topics were developed by retired US intelligence analysts.[4] When such topics are assigned to undergraduate research subjects, there may be a disparity in motivation and background knowledge. While it may not always be possible to create simulated work task situations, researchers should understand the limitations and constraints of whatever type of information needs are used.

8.2.3 Natural Tasks

Another type of information need that is used in IIR evaluations is natural information needs (or tasks). These are more commonly used in naturalistic studies where experimental control is less of an issue and the study focus is more exploratory in nature. Natural tasks are those tasks that subjects normally conduct in their everyday life. While the behaviors observed from subjects conducting natural tasks are more representative of those subjects' behaviors, it can be difficult to generalize and compare findings across subjects. Another difficulty is that tasks will likely be at varying levels of specificity, stability and completion, the amounts of information available to address different tasks will vary and subjects will know varying amounts about how to complete the tasks.

8.2.4 Multi-Tasking and Multiple Search Episodes

Most people acknowledge that users in the real world engage in multi-tasking and that information-seeking often takes place across multiple search episodes [181, 247, 251]. There have been a few studies of these behaviors, but there have not been a concentrated effort to develop tasks that would be appropriate for more controlled research settings. To study multi-tasking, one would need to create sets of tasks that can

[4] There are some collections where retired intelligence analysts or other US government personnel did not create the topics. For instance, topics in the TREC HARD 2004 collection were developed by summer interns at the Linguistic Data Consortium.

be done simultaneously. More importantly would be the development of a work task situation to provide background for why a user would want to conduct such tasks. The same can be said for tasks that need to be completed across multiple search episodes. Some research on re-finding has experimented with techniques for studying tasks that require at least two episodes (or search sessions) for resolution [48] and there is at least one example from IIR [181], but few researchers have attempted to study multi-tasking and multiple search episodes in IIR. There is not much guidance for creating tasks appropriate for these situations either, although Czerwinski et al.'s [66] diary study of multi-tasking behaviors in an office environment provides some groundwork.

9

Data Collection Techniques

Corpora, tasks, topics, and relevance assessments are some of the major instruments that allow researchers and subjects to exercise IIR systems. Other types of instruments, such as loggers, questionnaires and screen capture software allow researchers to collect data. This section provides an overview of some of the data collection techniques and instruments researchers have used to understand what happens during IIR evaluation.

9.1 Think-Aloud

The think-aloud method asks subjects to articulate their thinking and decision-making as they engage in IIR [87]. The researcher will need to somehow capture this data and this can be accomplished with an inexpensive computer microphone. Most computers come with recording software, so think-aloud data is relatively inexpensive to collect. However, there are problems associated with using the think-aloud method. Most subjects have a difficult time simultaneously articulating their thoughts and completing whatever IIR task has been assigned to them. In many cases, the IIR system is novel and subjects do not have the additional cognitive resources to engage in think-aloud. Think-aloud is

also awkward and unnatural — most people do not go around articulating their thoughts as they complete tasks. If subjects get quiet, the researcher can prompt them to continue thinking-aloud, but in general, people have a difficult time doing this and the researcher will likely do a lot of prompting, which might eventually annoy and distract subjects. Some researchers have proposed that subjects complete a short training task before they start searching to get accustomed to think-aloud. For instance, subjects might be asked to solve a puzzle while thinking-aloud before they start using an experimental system. Although this practice may help subjects become acquainted with think-aloud, it does not change its awkwardness. Finally, there is the question of whether the IIR task is too complex for think-aloud, since it was originally designed to be used with more basic tasks [87].

9.2 Stimulated Recall

An alternative to think-aloud protocol is stimulated recall (see [112] for an evaluation of this method and Rieh [217] for an application example). Stimulated recall (also called *talk after*) is used to collect the same type of data as think-aloud protocol, but differs in that data is collected during and after the search. The researcher records the screen of the computer as the subject completes a searching task. After the task is complete, the recording is played back to the subject who is asked to articulate thinking and decision-making as the recording is played. General instructions can be provided or the subject can be asked specific questions or to focus on specific features or processes. Although there is a delay between the subject executing the task and discussing what was going on, the researcher is generally able to get better data since the subject's attention is not divided. Discussing a completed event with another person is also a more natural activity than thinking-aloud during the event.

Instrumenting stimulated recall is more expensive than think-aloud since the researcher needs to purchase screen recording software. However, there are several relatively inexpensive pieces of software that can be used to record both the screen and audio. Such software can be used to record the screen while the subject searches and then to record the

screen and audio as the subject engages in stimulated recall. Screen recording software can also be used for other experimental tasks, such as creating tutorials.

9.3 Spontaneous and Prompted Self-Report

Another technique for collecting data from subjects while they engage in search is spontaneous or prompted self-report. This technique is not used a lot, in part because it is difficult to orchestrate and can be intrusive, but the basic idea is simple: elicit feedback from subjects periodically while they search. Subjects are not required to continuously verbalize their thoughts (as with think-aloud), but are instead asked to provide feedback at fixed intervals or when they think it is appropriate. The purpose of this technique is to get more refined feedback about the search that can be associated with particular events, rather than summative feedback at the end of the search. This is particularly useful in search situations that span long periods of time. Both qualitative and quantitative feedback can be gathered with spontaneous self-report.

Spontaneous self-report can be hard to integrate into a study because it is almost always intrusive, but one might imagine a persistent window that is visible on the screen at all times where subjects can communicate their feedback. Another method of implementation is to have a window pop-up at fixed intervals or have the researcher ask the subject to provide feedback at specific intervals. Either way, the subject is interrupted and might, after some time, become annoyed.

9.4 Observation

Observation can take two forms in IIR evaluation. In one form, a researcher is seated near subjects and observes them as they search or complete IIR activities. The researcher is usually trained to focus on particular events and behaviors and takes notes that describe their observations. Observation can also be conducted with a video camera or screen capture software. In these situations, the researcher usually conducts the observations at a later date, and the act of observing is actually combined with data analysis. Thus, observation can occur

in real-time or at play-back time (for recordings). Subjects might be uncomfortable because their actions are being monitored, but subjects are not required to do anything extra as with the previously described methods.

During real-time observation by a researcher, the subject is not interrupted, but can be asked follow-up questions about particular events later during post-search interviews. Typically researchers focus on recording things not available in a log, such as the subject's verbal or non-verbal reactions. Researchers might also note unusual interactions or instances when the subject seemed confused or frustrated. It is important to note that when observations are made at play-back time and the researcher has used screen capture software, it is not possible to record non-verbal communication. The only thing one can observe is what is happening on the screen. Non-verbal communication can be captured with a video camera, but recording people's faces puts them at more risk so one must be certain to take extra precautions with this type of data.

Observation is extremely time-consuming and labor-intensive, both in the collection and analysis of data. It is also prone to selective attention and researcher bias. It is extremely important that the researcher has some practice before commencing to observe and has some idea of what particular data need to be recorded and how it will be used. A protocol can be developed to guide the process, as well as a form to structure the observations. Training of observers is also important, especially if multiple researchers act as observers. Effort needs to be made to ensure that researchers record the same types of things and at the same level of detail.

9.5 Logging

One of the oldest and most common methods for collecting data in IIR evaluations is transaction logging (see [144] for a review). Papers were published at the *SIGIR Conference* in 1979 and 1983 describing the use of computer monitoring and its application to understanding user behavior [30, 77]. In 1993, there was a special issue of *Library Hi Tech* devoted to transaction log analysis [210], including an article on

its history and development [209]. The use of transaction logs has been around for quite some time, although the recent explosion of studies using Web transaction log data has re-popularized this approach. Four types of logging will be discussed: system, proxy, server, and client logging. Researchers face three major challenges in using log data: ensuring the validity and reliability of the logger, extracting and preparing data generated by the logger, and interpreting the data.

System logs refer to those that are written for an experimental IIR system. In this scenario, the logging application is built as part of the IIR system and records all interactions specified by researchers. The current practice is to write logs using XML, which greatly facilitates data extraction and analysis. System logs are typically used to characterize the interaction and record both what the system does as well as how the subject reacts. For instance, typical logs will record the subject's queries, the results shown to the subject and the results selected by the subject. This type of logging used to be referred to as transaction logging when applied to operational library systems. Because most current library systems are Web-based, there is a blurring between traditional system logs and Web logs, since in many cases these are the same things.

Common types of logging for studies conducted via the internet are server, proxy and client logging. The primary differences between these types of logging are where the logging takes place and what types of information are available to be logged. Server-side logging has traditionally been used in large-scale log studies conducted by search engine companies. Server-side logging takes place on a server and is thus limited to communications between the user and the server. Most servers come with some type of logging application to record basic data about which resources and services are requested and when they are requested. Such requests are typically associated with the IP address of the machine making the request as well as the client application that is making the request. One limitation of server-side logging is that only resources and services requested from a single server are recorded.[1]

[1] Currently, many search engine companies track users' activities after they leave the search engine server through a browser extension. While initial search logs from these companies

Thus, server-side logging does not track what subjects do once they leave a particular server. Server-side logging also does not record users' interactions with their local machines. This is a particularly important limitation since pages that are requested multiple times within a single period of time (for instance, when the user uses the back button on the browser) are not always routed through the server.

A final problem with server-side logging is that unless users are associated with particular usernames or IP addresses it can be difficult to determine unique identities and associate log records with these identities. In the past, the problem was primarily a result of different users using the same machine; before so many people owned computers, many shared computers that were found in a centralized location (e.g., computer laboratory and public library). Although more people own computers nowadays, many local networks are configured to dynamically assign IP addresses, which prevent one-to-one correspondences between users and IP addresses. Additionally, at least one browser extension has been created which corrupts the information that is sent to the server.[2]

Proxy logging also takes place at a server, but not the server from which the user requests resources. Rather, all communications between a user and a server are routed through an additional server, known as a proxy server. The proxy server may provide traditional internet services, but its main function is to log the communications that occur between a user and all servers to which they make requests. Proxy applications can also be used to modify or change what the user experiences. For instance, Joachims et al. [152] used a proxy application to change the order in which search results were displayed to subjects. Web browsers can be easily configured to point to a proxy server. While proxy logging allows the researcher to track the user's interactions with a number of servers, it still misses many actions that occur on the client machine.

The most comprehensive type of logging happens on a user's local machine via client-side application. Client-side logging is used to refer to logging applications that reside on the user's local machine. The

only contained information about communications with the server, more information is now available in these logs.

[2] http://mrl.nyu.edu/∼dhowe/trackmenot/

term *client* originally described particular applications that made requests to servers, but it is used as a synonym in this context to refer to the user's local machine (e.g., the user's desktop or laptop computer). Client-side logging applications can be used to record the user's interactions with all applications that run on the local machine, including Web browsers and word processing applications. Some of these loggers even record the user's interactions with the operating system. Client-side logging provides a more robust and comprehensive log of the user's interactions and solves most of the problems of server-side logging. Namely, unique identifiers can be associated with records if there is a one-to-one correspondence between the computer and the user and all requests, even those that operate on the cache, are recorded. Furthermore, additional activities that only occur on the client, such as scrolling and printing, are recorded.

Client-side logging is perhaps the most comprehensive type of logging, but it is also the most expensive and difficult to implement [89]. Building a client-side logger from scratch requires a great deal of knowledge and time. Glass box [61] is an example of a very robust client-side logger, but unfortunately it is only available to those working on certain US government funded projects. Jansen et al. [145] has made available Wrapper, an open source application designed to record subjects' Web interactions.[3] There are some commercial applications that can be used too, but researchers must be careful to evaluate these programs to make sure that the logs produce an accurate and reliable record of what occurred. Indeed, researchers bear this responsibility no matter what type of logging is used.

A final approach to logging involves instrumented Web browsers, such as those used by Kellar et al. [163]. This approach consists of creating a specialized browser or plug-in that works in conjunction with an existing browser. Instrumented browsers allow the researcher to focus on people's interactions with one particular application (i.e., the Web browser), to have more control over what is logged, and to log interactions with experimental features. Instrumented browsers also allow researchers to incorporate other data collection tools into

[3] http://ist.psu.edu/faculty_pages/jjansen/academic/wrapper.htm

the logger. For instance, the browser might contain widgets that allow subjects to classify the pages they view according to come criteria.

The primary benefit of any type of logging is that a record of the user's activities and interactions is created. The completeness of records varies depending on the type of logging (or more specifically where this logging takes place). Logging is also a useful method for capturing users' natural search behaviors in studies that occur outside laboratory settings. Most logging applications can run in the background while the user works without causing any disruption or delays. Perhaps the biggest limitation to logging is that only electronic observables can be captured — activities that occur beyond the computer or application are not captured and the purpose and intent of the observed actions is often unclear. Grimes et al. [111] conducted a comparison of data collected via query log, field study and using an instrumented browser and found that the query log provided the least useful data for individual events, but the most useful for understanding the scope of user's activities. Their paper title sums up their conclusion: query logs alone are not enough.

9.6 Questionnaires

The questionnaire is one of the most popular methods of collecting data from subjects in IIR evaluations. Almost all IIR evaluations have some type of questionnaire. Questionnaires can consist of closed questions where a specific response set is provided (e.g., a five-point scale) or open questions where subjects are able to respond in any way they see fit (e.g., what did you like most about this system?). Closed questions typically produce quantitative data, while open produce qualitative data. Thus, closed questions are useful for providing numeric representations of subjects' attitudes and feelings and allow researchers to make statistical comparisons. Open questions are useful for gaining more unique and varied insight into subjects' experiences and for understanding the reasons behind particular attitudes and behaviors. Responses to open questions also allow researchers to better interpret and contextualize subjects' responses to closed questions.

Closed questions typically take one of two forms: Likert-type scales or semantic differentials. In the first format, a series of statements are provided such as, *the system was easy to learn to use*, along with five- to seven-point Likert-type scales for responding, where one scale end-point (or anchor) represents strong agreement and the other represents strong disagreement. Traditional Likert scales measure agreement on a five-point scale, where the scale labels are: strongly agree, agree, neutral, disagree and strongly disagree [178]. The classic Likert measurement also produces a summative measure across a set of items. In IIR, a range of scales points are provided (although it is most common to have five- or seven-point scales), a range of scale labels are provided (e.g., not at all, somewhat, very much) and a summative measure is usually not produced, which is why this type of scaling is often referred to as Likert-type. There is little guidance about the appropriate number of scale points, although some researchers have suggested that seven-points are optimal for eliciting relevance assessments [265]. What is clear is that an odd number of scale points allow subjects to select a mid-point, while an even number does not. It is also the case that scales with more points can be converted into scales with fewer points.

The use of semantic differentials is another common way to format closed questions. Semantic differentials present pairs of antonyms at opposite ends of a scale without numeric labels. Instead of requiring subjects to identify discrete numbers, this scale allows subjects to place a mark along a line, indicating a more continuous-type of measure. While the semantic differential presents individual lines indicating where subjects should mark, other scales have been formatted to present a straight line without any demarcations. Of course, both of these formats only represent a difference in user interface: different points along the continuum are eventually coded into discrete numbers by the researcher for analysis.

Questionnaires can be administered via different modes: electronic, pen-and-paper and interview. Research in psychology and public opinion polling has found that questionnaire mode affects how people respond to questions [42, 274]. Kelly et al. [165] investigated questionnaire mode effects in the context of an IIR experiment and found that subjects' responses to closed-questions were significantly

more positive when elicited electronically, than via pen-and-paper or interview. Although this was only a single study with 51 subjects, these results suggest that questionnaire mode can impact subjects' response behaviors in IIR studies.

Questionnaires are used at various points during a study to collect data. The five most frequently used questionnaires are demographic, pre-task, post-task, post-system, and exit. The determination of which among this set is appropriate and what they contain depends on the purposes of the study. Table 9.1 lists some of the data that are typically elicited via each type of questionnaire.

In addition to these types of questionnaires, researchers sometimes give subjects specialized instruments (usually in the form of a questionnaire) that are used to characterize subjects along some standard measure (e.g., cognitive style, personality type, and spatial ability). These types of instruments are usually administered near the beginning of the study. It is also possible to use questionnaires to pre-screen potential subjects during recruitment if a particular characteristic is of interest. For instance, a researcher might only be interested in subjects who are novice searchers. Pre-screening can also helps to ensure a balanced design if a researcher is aiming for equal numbers of subjects across some characteristic (e.g., sex).

Instructions about how to design questions will not be presented here, but it is critical that appropriate attention is paid to this process and that questionnaires are piloted before they are used (see [207] for more information about question design). Important considerations to make when creating questions are related to the wording and ordering of the questions and choice of scale points, labels and anchors. With respect to wording, it is particularly important that the questions are not biased, loaded or double-barreled. For closed questions, it is also important that scale labels are appropriate to the question and that range of variation, exhaustivity, exclusivity and equivalence are considered. For open questions, the wording and placement of questions within the questionnaire are critical issues.

There are additional considerations to make about closed and open questions. Because a response set is provided for closed questions, the researcher has to ensure that the response set is appropriate. There is

Table 9.1 Commonly used questionnaires in IIR evaluations.

Questionnaire	Purposes	Administration
Demographic	This questionnaire is used to elicit background information about subjects. This information is typically used to characterize and describe subjects, but it can also be used to explore and test specific hypotheses. For instance, a researcher might be interested in investigating the difference between male and female behavior, or among people with different amounts of search experience.	This questionnaire is usually given at the start of the study, but it can be given at the end. The rationale for waiting until the end is that subjects are likely to be fatigued and it is better to get this "easy" information then.
Pre-task	This questionnaire can be used to assess subjects' knowledge of the search task and/or topic. Questionnaire items are usually directly related to the search task in which the subject is about to engage.	Subjects complete this questionnaire before searching occurs so that the search experience does not bias responses.
Post-task	This questionnaire is most often used to gather feedback about the subject's experiences using a particular system to complete a particular task. Thus, the primary goal of this questionnaire is to assess the system–task interaction.	This questionnaire is administered following each task.
Post-system	This questionnaire elicits feedback from subjects about their experiences using a particular experimental system. It is typical to administer this type of questionnaire during within-subjects studies where subjects use more than one system. The assessment is usually focused on the subjects' experiences using the system to complete a number of tasks and represents an overall assessment of a particular system.	The questionnaire is administered after subjects finish using a system. Subjects complete one questionnaire for each system.
Exit	If the study is a between-subjects study, then this questionnaire functions similarly to the post-system questionnaire. However, for within-subjects studies, this questionnaire can be used to elicit cross-system comparisons and ratings.	As its name implies, it is typically administered at the end of the study.

a danger that the questions and response sets will bias subjects since appropriate topics and responses are suggested. Data elicited by closed questions is homogenous and easier to analyze, but it is important to remember that the numeric scales are not based on a true number line, but instead represent a set of labels (albeit sometimes ordered). They are subject to individual interpretation and can be difficult to compare; there is no true zero and one subject's six may be another subject's five. Open questions take longer to administer than closed questions and responses are more difficult to interpret and analyze. People typically use different words to describe the same things and some subjects are better at clarifying and explaining their responses than others.

9.7 Interviews

Few IIR evaluations consist solely of interviews, but interviews are a common component of many study protocols. Most often, the interview is used as a delivery mode for a set of open-ended questions which might just as easily be delivered via print or electronic questionnaire. Although an interview mode is not necessarily required to ask such questions, it presumably allows one to get more individualized responses and allows some flexibility with respect to probing and follow-up. Kelly et al. [165] compared subjects' responses to a set of open-ended questions across three modes: interview, pen-and-paper, and electronic and found that while subjects' responses were longer in the interview mode than in the other two modes, the number of unique content-bearing statements they made in each mode were about equal. These results do not mean that the interview mode is not useful, but rather that it may not be useful for certain types of questions. The researchers asked questions that were similar to those asked in traditional IIR evaluations, which really are not designed with depth-interviewing in mind. If one were interested in asking more complex, abstract questions then it is likely that the interview mode would be more appropriate.

Another place where interview techniques can be used in IIR evaluations is during stimulated recall. During stimulated recall, subjects verbalize their decision-making processes and thoughts while watching

a video recording of a search they recently completed. During this process, the researcher might interrupt with specific pre-planned questions. These questions might be used to find out about something specific, or to probe remarks or actions made by subjects.

When planning to conduct interviews, researchers must first decide what type of interview they would like to administer: structured, semi-structured or open. This will be dependent on the purpose of the research. Researchers typically create an interview schedule which is a list of all of the questions they would like to discuss with the subject. In some cases, the researcher might go through this list of questions one-by-one in the same order for all subjects. This is typically the case with the short interviews that are conducted at the end of an IIR study. When the interview is the primary method used in a study, it is common to skip around and diverge from the list of questions. In these types of interview situations, the researcher has more flexibility and subjects, in many ways, direct the interviews since they can determine what topics will be discussed and in what order. The interview schedule functions to remind researchers of all of the topics that they would like to cover, but it usually does not determine the sequence of questioning. Along with an interview schedule, a researcher might also use other types of stimuli to facilitate and structure the interview. For instance, screen shots of an interface can be used to provide a focal point for discussion or even a video recording of a confederate using the experimental system.

9.8 Evaluation of End Products

A final method of collecting data during IIR evaluations focuses on the outcome or product of the search. Very often the focus is on the resolution of the work task, as opposed to the search task [in Byström and Hansen's [44] language]. This approach is used less frequently because it is more time consuming for subjects since it requires them to complete an additional and more complex task beyond finding relevant documents. The additional task is to actually use the information in a way that matches the user model behind the search task. For instance, the user model behind the traditional IR search task is a

person collecting materials that can be used to generate a report about a particular topic. The methods discussed in this section ask subjects to use the documents they find to create papers, reports, or other end products. These end products, in turn, are studied along with the subject's interactions with the system. Example studies that have used such methods include Egan et al. [83], Halttunen and Järvelin [116], Kelly et al. [166], Marchionini and Crane [192] and Vakkari [280].

From a philosophical point-of-view, this approach differs from others discussed in this section in that what is being evaluated and studied extends beyond the system. The IR system is viewed as a tool that helps people accomplish some other goal; thus, IR is not viewed as an end unto itself, but as an activity that supports a larger goal. The approaches discussed here focus on examining that larger goal. The underlying notion is that a better IR system will help people do a better job of achieving this goal (e.g., write better quality reports).[4] These approaches also assume a different perspective of relevance. Specifically, a distinction is made between the differences in relevance behavior exhibited during information finding and information use. The notion is that subjects engage in a different kind of relevance behavior when they are selecting relevant documents than when they are making choices about what to actually use and include in the final product.

The methods described in this sub-section are all executed within the context of a traditional IR task model where people engage with an IR system to find documents that support the writing of a report. Some researchers have studied students who are working on a writing assignment for a course, while others have included report writing as part of the study protocol. The former approach requires coordination with an instructor and it may not be possible to use an experimental IR system for such critical tasks. The latter requires the researcher to create a writing task, which might be artificial since the subjects are not performing it for any purpose other than the study. The difficultly in both cases is that many things contribute to the end product, including a person's writing and organizational skills. However, if two or more

[4] It may become increasingly difficult to analyze the system's contributions as subjects are able to compensate for poorly performing systems [243].

groups are being studied (e.g., if there are two systems being studied), then randomly assigning subjects to groups should distribute subjects with differing skills equally.

IIR researchers have used several approaches to study end products: examination of references, expert assessments and cross-evaluation. Examination of references is the most basic way to evaluate the extent to which the documents found by the user during searching were used to complete the larger task. The reference list, of course, does not tell the entire story. It is likely that many documents used during other information-seeking stages do not make it into the reference list. These documents play an important role in the creation of the final product, but they are not visible from the reference list.

Researchers have also asked experts to assess the quality of the final products [280, 282]. If the study is conducted in conjunction with a class, then this expert might be the course instructor who assigns grades to the final products. Course instructors will be experienced evaluators of the end products and will likely have established grading rubrics (even if they are internalized). If the evaluators are not very experienced, then the researcher will need to develop a rubric and spend time training evaluators to use it consistently. While these final products are a more realistic representation of how the information is used, it is difficult to coordinate and administer this type of assessment.

Cross-evaluation was developed as both a method and tool to facilitate comparison and rating of reports generated by subjects of information systems [257]. It requires subjects to perform three activities: use an information system to find relevant information, create a short report summarizing their findings, and evaluate other subjects' reports. Reports are evaluated according to seven quality criteria related to aspects of the information contained within the reports as well as aspects of the report itself. Cross-evaluation was developed in the context of interactive question–answering systems with intelligence analysts as subjects, but it could be extended to other types of situations. Interestingly, cross-evaluation can be a motivator for subjects since their work will be reviewed by others.

10

Measures

Measurement is fundamental to IIR research, but there are few research programs dedicated exclusively to the development and evaluation of measures for IIR. A large number of measures have been used but, at least until 1992, most could be categorized as relevance measures, efficiency measures, utility measures, user satisfaction, and success measures [254]. Su (2003) conducted a second review of evaluation measures in 2003 in preparation for an evaluation of search engines. Similar classes of measures were found except that success was replaced with connectivity. In her review of IIR research, Su (2003) further identified the following classes of measures for characterizing subjects and their information needs and behaviors: background (e.g., professional field, age, and sex); experience (e.g., use of IR systems and use of the internet); and information needs/search requirements (e.g., search topic, purpose of search, and time period of documents). Yuan and Meadow [299] also reviewed the research and created a classification of variables used in IR user studies. Classes included variables related to the study participants, searches and outcomes. Variables related to searches included an extensive list of items that ranged from specific

search tactics to performance measures. Boyce et al. [39] also compiled a list of measures in information science research.

Over time, four basic classes of measures have emerged as the standard: contextual, interaction, performance, and usability. The first set of measures includes those used to characterize subjects: such as age, sex, search experience, personality-type, and those used to characterize the information-seeking situation: such as task-type and subjects' familiarities with topics. Also included in these measures are geographic location and time. These measures basically describe the context in which information search occurs. It is beyond the scope of this paper to list every possible measure that fits this description. Ingwersen and Järvelin [139] provide an extensive discussion of context in information seeking and retrieval, while Dourish [78] provides a theoretical examination of the concept. Contextual measures in IIR evaluations can be elicited via questionnaires (e.g., age, sex, and topic familiarity) or controlled by the researcher (e.g, task-type). Many of these measures can be characterized as socio-cognitive measures or individual difference measures. While researchers have discussed context for many years and the difficulties with defining and measuring it (e.g., [6, 58, 75]), recently there have been large efforts in IR and IIR to systematically incorporate context into retrieval and evaluation (see e.g., [38, 227]).

The second set of measures includes those used to characterize the interaction between the user and the system and the user's search behaviors, such as number of queries issued, number of documents viewed and query length. These types of measures are typically extracted from log data.

The third set of measures are performance-based measures related to the outcome of the interaction, such as number of relevant documents saved, mean average precision, and discounted cumulated gain. These measures are also typically computed from log data.

The final set of measures includes those based on evaluative feedback elicited from subjects. Such measures often probe subjects about their attitudes and feelings about the system and their interactions with it. Although this class of measures is referred to as *usability* for simplicity sake, this class of measures includes a variety of self-report measures.

In IIR, performance measures have typically been separated from usability measures, even though as we will see shortly, *effectiveness* and *efficiency* are standard dimensions of usability and are often measured in HCI and ergonomics research with measures such as recall, task completion, error rate and time. Thus, in HCI performance is most often subsumed under usability, while in IIR we treat performance as a separate entity from usability and usually use the term *usability* as a synonym for self-report measures. The likely reason for our unusual use of this term is that performance measures have always been part of IR evaluation even before it was common to include subjects in evaluations. As it became more common to study subjects, various self-report-based usability measures made their way into the evaluation literature, but performance was still a first-class measure. Many early IIR evaluations discussed measures such as satisfaction and user-friendliness without mentioning the term usability. Common dimensions of usability are *effectiveness*, *efficiency*, and *satisfaction*, and what typically happens in IIR is that self-report measures are used to elicit evaluative data from subjects about these qualities. Again, this is contrary to how usability is measured in HCI — effectiveness and efficiency measures often consist of objective measures (e.g., some of our standard performance measures), while satisfaction measures are elicited via self-report.

Regardless of how one labels such measures, the most important thing is to be clear about the definitions of the measures. Devising appropriate measures involves the provision to two basic types of definitions: *nominal* and *operational*, which were discussed in a previous section. Nominal definitions state the meaning of concepts; for example, one might define performance as the ability to find relevant documents. Operational definitions specify precisely how a concept (and its dimensions) will be measured. For instance, an operational definition of *learnability* might include three questions and a five-point Likert-type scale for responding. Alternatively, one might operationalize learnability as the length of time it takes a user to learn to use an interface. Without both nominal and operational definitions it is impossible to know exactly what concepts researchers hope to capture with their measures, and it impossible to evaluate the credibility of the results.

There are also many validity and reliability considerations that are related to measurement, which are discussed in Section 12. Historically, IIR researchers have not been as concerned about measurement validity and reliability as other researchers who rely on self-report measures and questionnaires to elicit data from human subjects. Thus, studies that use self-report metrics which have not undergone serious scrutiny with respect to validity and reliability are susceptible to measurement error caused by the items themselves. Measurement bias is well-documented in many literatures including psychology, public opinion polling and public health. There is an entire field called psychometrics, which is devoted to understanding how to better measure psychological phenomena. This research has shown that people exhibit a number response biases when completing self-report measures, including inflation where the tendency is to rate things more positively than they are, and acquiescence where the tendency is to agree with everything.

Studies have also shown that slight variations in study procedures can impact study results and that method variance, in general, is almost always a potential problem [215]. Method variance refers to variance that is attributable to the measurement method rather than to experimental stimuli [94]. Systematic error variance, of the kind that might be caused by invalid or unreliable measurement techniques, is particularly problematic since it can produce results that might appear meaningful, but are only a function of the measurement technique. Method variance is a threat to any study regardless of whether self-report data are being elicited, but it is a more acute problem for studies that rely on self-report data. It is also problematic when no serious attempt is made to generate valid and reliable measurement techniques and the accepted standard is to just generate measures, especially self-reported ones, in an *ad-hoc* fashion.

Finally, the selection and interpretation of any measure, and particularly performance measures, should be grounded by the purposes of the system and the task the user is trying to accomplish. If a user is asked to complete a high-precision task such as finding one or a small number of documents that answer a particular question, then assessing recall makes little sense, since it is not appropriate to the retrieval

situation. If a system is designed to support exploratory search, then more interaction might be better than less.

10.1 Contextual

The measures presented in this section describe the context in which information search and interaction occurs. These include measures used to characterize subjects, such as age and sex, and those used to characterize the information need, such as task-type and domain expertise. It is common in most IIR evaluations to elicit some basic measures to describe study subjects, tasks and information search situations, but these measures are not always used as independent variables.

10.1.1 Individual Differences

Boyce et al. [39, p. 202] state, "the purpose of measuring user characteristics separately from the search process is to be able to use them to predict performance or to explain differences in performance." Such differences are often referred to as individual differences. Borgman [31] and Dillon [76] provide overviews of individual difference research. Borgman's article is focused on individual differences in information retrieval. Dillon's article is targeted to HCI researchers and strongly grounds individual differences in psychology research. Individual differences research was very popular in IIR during the period 1980s to 1990s, but has not received as much attention lately.

Fenichel [88] provides an overview of some of the more common measures of individual differences in the context of online searching. These include variables such as sex of subject, age, college major, profession, level of computer experience, and level of search experience. The latter two variables, in particular, do not figure as prominently in current research because there is not as much variability in these factors as there used to be, especially considering the typical subject in most IIR evaluations. Today, if one wanted to study many of these variables, one would need to purposively sample for them. However, there are still expected differences between many groups of people — for instance, one would expect subjects with advanced degrees in library science, computer science or human–computer interaction to be different from

subjects from the general population [90, 237, 238]. Therefore, it is important to report such characteristics, even if they are not independent variables, and carefully consider how they might impact the study results. Morahan-Martin [198] reviews research related to sex differences and internet usage, while Ford et al. [100] and Lorigo et al. [185] investigate sex differences related to information search. Ford et al. [99] investigate internet perceptions and cognitive complexity as additional ways to measure individual differences.

Another set of individual difference measures are those related to intelligence, creativity, personality, memory, and cognitive style. One nice thing about studying these types of measures is that there are a large number of standardized instruments found in the education and psychology literatures. Cognitive style, in particular, continues to receive a great deal of attention. Cognitive style is related to how people think about and approach problems. Ford et al. [100, 99], citing Riding and Cheema [216], state that there are two basic aspects of cognitive style: the wholist-analytic style characterizes users according to whether they tend to process information in wholes or parts and the verbal imagery style characterizes users according to whether they are verbal or spatial learners. Example instruments for assessing cognitive style include the learning style inventory, the remote associates test, the symbolic reasoning test and the Myers Briggs type indicator (Borgman, 1987). Ford et al. [99] use the cognitive styles analysis and approaches to studying inventory (which categorizes people as engaged in deep learning, surface learning or strategic approach) to measure cognitive style. The effects of cognitive style on information-seeking behavior in mediated search situations have also been investigated by Ford et al. [101]. Finally, computer and search self-efficacy have been studied as more refined measures of computer and search competency [57, 71].

10.1.2 Information Needs

Another important set of contextual variables are those that characterize the information need. Example measures include those related to the task such as task-type, task familiarity, task difficulty and

complexity, and those related to the topic such as a topic familiarity and domain expertise.

One problem with studying many of these items is that it is difficult to devise instruments for measuring them. For instance, topic familiarity is often measured with a seven-point scale, which does not really provide much information about how much a person knows about a topic. Since such scales are not calibrated it is even more difficult to make comparisons across familiarity levels. Domain expertise is often measured using credentials — for instance, a person with an advanced degree in molecular biology might be said to have high domain expertise in this subject area. Again, such coarse classifications often make it difficult to interpret study results and reach conclusions.

Other attributes of the information need that are often measured include persistence of information need, immediacy of information need, information-seeking stage, and purpose, goals and expected use of the results. Ingwersen and Järvelin [139] provide an extensive review of many of these types of variables.

10.2 Interaction

Interaction measures describe the activities and processes that subjects engage in during IIR. These measures are basically low-level behavioral data — such behaviors might originate from the subject or the system. Some types of interaction measures are common to almost all IIR evaluations. For example, number of queries, number of search results viewed, number of documents viewed, number of documents saved, and query length. Other types of interaction measures are specific to the individual system being studied. Many interaction measures are frequency counts of the activities that occurred and can be related directly to interface functionality. This includes basic *uptake* measures that show with what frequency subjects are using a feature or application.

Since most interaction measures are counts, they are continuous data types and can be combined to form other measures. For instance, time can be divided by the number of documents saved, or the number of documents saved can be divided by the number of documents

viewed. Usage patterns can also be derived from interaction measures. For instance, Markov modeling can be used to determine the most probable sequences of actions or a researcher can attempt to assemble low-level interactions into search tactics and strategies [15, 90].

One of the most challenging aspects of using interaction measures is developing a framework for interpreting them. Relating these signals to concepts requires one to consider the purpose and nature of what is being studied. If a subject enters a large number of queries, is this good or bad? The answer to this question is likely related to the purpose of the system — if the purpose of the system is to help a subject learn more about a topic, then more queries might be a positive indicator. If the purpose of the system is to help a subject find a single answer, then more queries might be a negative indicator.

Thinking more broadly about interaction, another important question to ask is what is interaction? Interaction with computers has been studied in a number of disciplines and there is not necessarily any agreement over what it means. There has been few serious theoretical discussion of the concept of interaction in IIR (e.g., [16]). In IIR, an implicit definition of interaction is accepted, which is very closely tied to feedback. Spink [248] and Spink and Losee [250] provide extensive discussions of the nature of feedback and identify interactive feedback units, the smallest of which consists of the user responding to the system and the system using the user's response to produce new content.

10.3 Performance

While there are many well-established measures for classic IR performance evaluation, there are not too many for IIR. As a result, many IR measures are used in IIR evaluations. The fit is not always perfect and it is probably safe to say that most researchers are not completely satisfied with these measures. Nested within most of these performance measures is another measure — relevance, and this is where things often break in an IIR evaluation scenario. The major problem is that most classic IR measures were developed and evaluated under different retrieval circumstances where certain assumptions could be made about relevance judgments and behaviors. These assumptions

underlie many of the evaluation measures developed as part of TREC, so the applicability and usefulness of these measures to IIR evaluations can be questioned. When using a TREC collection in an IIR study, researchers must assume for the purposes of the study that relevance is binary (usually), static, uni-dimensional and generalizable. Although these assumptions are incongruent with a lot research and with what most researchers believe, it is a *suspension of disbelief* that is required to use the TREC collection. Although there have been some attempts to create test collections with graded relevance assessments (e.g., [245]), this has been the exception rather than the rule. As mentioned previously, most standard IR performance measures assume binary relevance and do not easily accommodate situations where there are graded relevance assessments.

There have been a series of studies that have compared results of batch-mode and user studies and found that systems which perform better in batch-mode studies do not always do so in user studies [129, 243, 275, 276]. There are a number of explanations of these findings and many relate to the nature of relevance. Specifically, users often discard documents that TREC assessors have found relevant and find and save documents that TREC assessors never evaluate. Since it is the TREC assessor's judgments that are used to evaluate system performance, conflicting results are possible depending on how performance is evaluated. Another issue to consider when using TREC-based performance metrics in IIR evaluations is whether the metric is actually meaningful to real users. A measure that evaluates systems based on the retrieval of 1000 documents is unlikely to be meaningful to users since most users will not look through 1000 documents. Furthermore, it is important to note the limitations of performance measures: these measures can show that a system is functional (if used in a systems-centered evaluation), but not necessarily usable.

Finally, the actual technique used to measure relevance can vary considerably. There have been a number of studies that have examined and studied (1) the concept of relevance (e.g., [33, 197, 224, 232, 235, 236]); (2) the criteria users employ when making relevance assessments (e.g., [271]); and (3) techniques for measuring relevance (e.g., [84, 162, 272, 283]). Suffice to say, the published research about how users make

relevance assessments and the actual measures that researchers employ to collect relevance assessments are not very aligned.

10.3.1 Traditional IR Performance Measures

Table 10.1 presents some of the classic IR performance measures. For more information about these measures and descriptions of other IR measures, see Voorhees and Harman [288].

10.3.2 Interactive Recall and Precision

The measures in Table 10.1 are based on an assessor's relevance judgments and batch-mode retrieval runs that can consist of up to 1,000 documents. In IIR evaluations, subjects usually are unable to search through 1,000 documents. It is also the case that subjects are typically instructed to save documents that they find relevant and as described earlier subjects' relevance judgments often do not agree with the assessor's relevance judgments, so using the benchmark relevance judgments to assess performance may not be meaningful. Some TREC topics have hundreds of documents that have been marked as relevant by assessors and it is unlikely in most situations that a subject will search long enough to find all of these documents.

To partially account for the mismatch between TREC relevance judgments and subjects' relevance judgments, Veerasamy and Belkin [284] and Veerasamy and Heikes [285] proposed the use of interactive recall and precision, and interactive TREC precision which compares TREC relevant documents with those saved by subjects (Table 10.2). *TREC relevant* means the document was marked relevant by an assessor. These measures basically try to account for the two-stage relevance process that happens in IIR evaluations that use collections with relevance judgments: first, an assessor makes a relevance judgment and then a subject makes a relevance judgment. Documents that the assessor marks as relevant may or may not be retrieved, viewed or saved by subjects.

Relative relevance (RR) is a measure for comparing the degree of agreement between two relevance assessments [33, 36]. This might be between the system's relevance score and a subject's or between an

Table 10.1 Some classic IR evaluation measures.

Measure	Description
Recall	The number of retrieved relevant documents divided by the number of relevant documents in corpus.
Precision	The number of relevant retrieved documents divided by the number of retrieved documents.
F-measure	The F-measure is a way of combining precision and recall and is equal to their weighted harmonic mean $[F = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})]$. The F-measure also accommodates weighting of precision or recall, to indicate importance.
Average precision (AP)	Individual precision scores are computed for each relevant retrieved document (with 0 assigned to relevant documents that are not retrieved). These values are then summed and divided by the total number of relevant documents in the collection. Thus, AP has a recall component to it and is typically described as the area underneath the precision/recall curve. AP also takes into account the position of relevant documents in the result list.
Mean average precision (MAP)	This is a run level measure and consists of taking the average of the average precision values for each topic.
Geometric average precision (GMAP)	The geometric mean of n values is the nth root of the product of the n values. Robertson [219] recommends taking the logs of the values and then averaging. GMAP was developed for the TREC Robust Track, which explored retrieval for difficult topics and does a better job than MAP of distinguishing performance scores at the low end of the AP scale.
Precision at n	The number of relevant documents in the top n results divided by n. Typical values for n are 10 and 20, which is thought to better represent the user's experience since research has shown that this is the extent to which users look through Web search results [146].
Mean reciprocal rank (MRR)	This measure was developed for high-precision tasks where only one or a small number of relevant documents are needed. For a single task with one relevant document, reciprocal rank is the inverse of its ranked position. MRR is the average of two or more reciprocal rank scores (used when there is more than one task).

assessor's relevance scores and a subject's. Thus, this measure attempts to accommodate the subjective nature of relevance and provide a measure of the extent of this overlap. RR might also be used to evaluate the extent to which assigned information need descriptions are

Table 10.2 Modified versions of recall and precision for interactive IR [284, 285] and relative relevance [33, 36].

Measure	Description
Interactive recall	Number of TREC relevant saved by user/number of TREC relevant documents in the corpus.
Interactive TREC precision	Number of TREC relevant documents viewed by the user/total number viewed.
Interactive user precision	Number of TREC relevant documents saved by the user/total number saved by the user.
Relative relevance (RR)	Cosine similarity measure between two lists of relevance assessments for the same documents.

well- or ill-defined — presumably there will be less overlap in relevance judgments for ill-defined needs.

10.3.3 Measures that Accommodate Multi-Level Relevance and Rank

Two other problems with traditional performance measures is that they only accommodate binary relevance assessments and they do not take into account that relevant documents that are retrieved further down on the results list are less useful since subjects are less likely to view them. Not only must a subject expend some effort to get to these documents, by the time the subject arrives at the document its content may be less valuable because of what the subject has learned on the way to the document. Traditional measures such as precision and recall do not accommodate this situation. Two results lists — A and B — may have the same recall and precision scores, despite having very different orderings of the documents. MAP was created to address the ordering problem in systems-centered research, but it still maintained some of the problematic assumptions of the traditional TREC measures.

Järvelin and Kekäläinen's [148, 149] suite of cumulated gain measures and Borlund and Ingwersen's [36] ranked half-life measures are measures that have been created for use in interactive search situations where human searchers make relevance judgments (Table 10.3). In addition to these measures, Borlund [33] identifies several others that account for position of relevant documents including Cooper's

Table 10.3 Cumulated gain measures [148, 149] and ranked half-life [33, 36].

Measure	Description
Cumulated gain (CG)	Cumulated gain can be computed at different cut-off values for search result of lists of varying sizes. At the cut-off point, CG is the sum of the relevance values of all documents up to and including the document at the cut-off point.
Discounted cumulated gain (DCG)	Discounted cumulated gain discounts the value of relevant documents according to their ranked position. New relevance values are computed by dividing the relevance score of a document by the logarithm of its rank. The discounted relevance scores are then summed to a particular cut-off point.
Normalized discounted cumulated gain (nDCG)	The DCG measure is normalized according to the best DCG available for a given results list. This normalization transforms DCG scores, which can take on a large range of numbers, to a 0–1 scale, which is easier to interpret and compare.
Ranked half-life (RHL)	The point in the results list at which half of the total relevance value for the entire list of documents has been achieved. If binary assessments are used, this is the point at which half of the relevant documents in the list have been observed. If multi-level assessments are used, this point is when half of the sum total of all of the relevance values are observed.

[59] *expected search length*, Dunlop's [81] *expected search duration* and Losee's [186] *average search length*. Käki and Aula [158] also propose *immediate accuracy* which is the proportion of cases where subjects have found at least one relevant result by a particular cut-off position in a ranked list of results. Kekäläinen and Järvelin [162] also extend traditional precision and recall metrics to accommodate graded assessments.

Discounted cumulated gain is based on the notion that the lower a document's rank in a results list, the less likely the subject is to view it. For instance, the chances of a subject viewing a document ranked in position one in a search results list is greater than the chances of the subject viewing a document ranked 88th. The measure also assumes that the number of topically relevant documents in a corpus is likely to exceed the number of documents a subject is willing to examine [162]. This measure also allows for multi-level (or graded) relevance assessments, which makes it more versatile and reflective of how most subjects make relevance assessments. To compute cumulated gain the

search results are first viewed as a vector where each document is represented by its relevance value. The cumulated gain of each document is a function of its relevance value plus all of the relevance values of the documents ranked above it.

To compute discounted cumulated gain, the relevance value of a particular document is treated as a function of a document's relevance and its rank. The *discounted* part of the measure reduces the contributions of relevant documents that are ranked lower in the list. This is accomplished by dividing the relevance of the document by the log (base 2) of its rank.[1] For instance, assuming four levels of relevance, where four represents a highly relevant document, a highly relevant document at rank 16 would contribute a score of 1 [$4/\log_2(16)$, or $4/4$]. The base of the log can be adjusted to match varying types of users. For instance, patient or impatient users. DCG scores can also be normalized (nDCG) based on an ideal result list which can be created by ordering all documents judged from most relevant to least relevant. Cumulated gain-based measures are typically reported at particular cut-off values. In Kekäläinen and Järvelin [162] the use of graded relevance assessments (non-binary based) are extended to a number of other evaluation measures as well. Järvelin et al. [150] extended DCG for use in multi-query sessions.

Rank half-life (RHL) measures the extent to which relevant documents are located at the top of a ranked results list [33, 36]. This measure is similar to MAP and DCG in that not only is the relevance values of particular documents included in its calculation, but also their position in a ranked list. There are two measures associated with RHL — RHL and RHL-index. The RHL is the position at which half of the relevant documents are retrieved. If multi-level relevance values are used, this is the point at which half of the total relevance value for the entire list of results has been observed. The formula used to calculate RHL is the basic formula for the median of continuous data. Thus, lower RHL are associated with better retrieval performance since lower numbers indicate that more of the relevant documents were found

[1] The discount is not applied to the document in position one of the search results list since the logarithm of this would result in a denominator of zero.

near the top of the results list. The RHL-index allows one to compare two result lists at a particular cut-off value, given a precision value. The RHL-index is the RHL of the list divided by the precision of the list.

10.3.4 Time-Based Measures

Time has been used quite a lot in IIR evaluations, both at a gross level (e.g., the length of time it takes a subject to complete a search task) and a more specific level (e.g., the length of time a subject spends viewing a search result or engaging in a specific action). As mentioned previously, time-based measures can be difficult to interpret since this is dependent on the task, the objective of the system and the researcher's beliefs about IIR. Time-based measures are often used as indicators of efficiency, although as stated earlier, effectiveness (performance), efficiency and satisfaction can be separated from one another. Efficiency and time will be discussed again in Section 10.4 when traditional notions of usability are discussed along with evaluative self-report measures.

Researchers have used a variety of time-based measures, including the length of time subjects spend in different states or modes, the amount of time it takes a subject to save the first relevant article, and the number of relevant documents saved during a fixed period of time. The number of actions or steps taken to complete a task is another way to look at time and efficiency. Käki and Aula [158] formalize two time-based measures that have been used in IIR research, search speed and qualified search speed (Table 10.4). These measures are based on answers not relevant documents, but could be extended to cover this retrieval unit.

Table 10.4 Time-based measures from Käki and Aula [158].

Measure	Description
Search speed	The proportion of answers that are found per minute. This measure consists of dividing the total number of answers found by the length of time it took to find the answers. All answers are included in this computation regardless of whether they are correct.
Qualified search speed	This measure accommodates multi-level relevance and consists of computing search speed for each relevance category, including non-relevant.

Although it is more common to consider how long it takes subjects to perform particular actions, the length of time it takes the machine to perform particular actions is also a common time-based efficiency measure in IR. Most would agree that this impacts a subjects' experience with a system and likely contributes to their evaluation of the system. Cleverdon et al. [55] discuss the response time of the system, which can be measured with a simple time-based figure, or could be an analysis of computational complexity. Common measures of computational complexity are number of steps, iterations or computing cycles that are needed by the computer to perform a task and the amount of computing resources needed.

10.3.5 Informativeness

Informativeness is a measure of the output of a system proposed by Tague [258, 261, 263]. This proposed method for evaluating search results focuses on relative evaluations of relevance rather than absolute measures. The assumption behind this is that asking subjects to rank a set of search results from most informative to least informative results in more accurate data than asking them to associate absolute judgments with each result using a scale. While Tague [263] wrote quite a bit about the informativeness measures and explored this measure in the context of browsing, a large-scale validation of this measure was never achieved due to her death [103]. Freund and Toms [103] recently re-introduced this measure and explored it in the context of Web search. Interestingly, there are many current proposals to use relative relevance judgments to evaluate search results lists (e.g., it is generally accepted that clicks equal relevance (whether right or wrong)). Perhaps with this renewed interest in relative relevance judgments, Tague's informativeness measure will finally be validated and adopted as a standard method of evaluation.

10.3.6 Cost and Utility Measures

In the early days of IIR research, cost and utility measures figured prominently in the IR evaluation framework. Some researchers treated these measures as separate constructs from relevance, while

others attempted to use these measures as substitutes. Cooper [60] proposed the use of subjective utility as the benchmark with which systems should be evaluated. In his proposal, users associated dollar amounts with search results. Salton [230, p. 442] summarizes the utility-theoretic paradigm, "retrieval effectiveness is measured by determining the utility to the users of the documents retrieved in answer to a user query". Salton [229] identifies a host of cost-based measures including those associated with the operational environment, response and processor time.

Belkin and Vickery [23] identified utility as one of the major approaches underlying performance measures alongside relevance and user satisfaction. In a study of evaluation measures for IIR, Su [254] compared 20 measures, including actual cost of search and several utility measures such as worth of search results in dollars, worth of search results versus time expended and value of search results as a whole. Su [254] found the value of the search results as a whole was the best single measure of IIR performance. The 40 subjects involved in this study were responsible for the costs of their own searches which likely changed the importance of this variable to them.

The popularity of utility measures in the early years is not surprising since users were charged to use many operational IR services. Utility and cost functions have always been an important part of the evaluation of library and information services (e.g., [239]). Even though people are still charged to access databases and view the full-text of articles, this cost is usually incurred by the user's institution, thus the price of information services are often out of users' awareness, despite continuing to be an important issue for institutions. Furthermore, since so much information is freely available online these types of measures are arguably less relevant to the individual user. It is likely the case that when using IIR systems, most users are not thinking about the costs associated with the service or information, at least not in terms of monetary values.

Recently, Lopatovska and Mokros [183] investigated willingness to pay and experienced utility as potential measures of the affective value of a set of nine Web documents. While the results are limited given the small number of documents evaluated, subjects' responses seemed to

suggest that willingness to pay reflected the rational value of the documents for completing the task, while experienced utility reflected an emotional, task-neutral reaction to the documents. This work suggests that these measures may continue to have value in IIR evaluations.

10.4 Evaluative Feedback from Subjects

Much of the data elicited during an IIR study is self-report, evaluative feedback from subjects. Very often researchers refer to these as *usability* measures, but this is not entirely appropriate. In many cases, researchers do not provide any conceptual or operational definitions of usability and instead lump all self-report data together and call it usability data. Referring to all self-report data as usability data overly-restricts the types of questions that can be asked and does not encourage much thought about the nature of the data that is collected. Also, as discussed earlier, traditional usability measures typically include objective performance measures, but in IIR, performance has been treated as a separate category because of its importance in classic IR.

HCI research has shown that there is only a slight correlation between objective and subjective (i.e., self-reported) performance metrics, and that people tend to use the high-end of the scale when evaluating systems (i.e., inflation) [133, 203]. Anecdotal evidence from IIR research also suggests this, and recently researchers in IIR are starting to look at this empirically [168]. Whether or not one believes that objective and subjective measures should be correlated is a theoretical issue that has not been explored in IIR. While it can be argued that response biases such as inflation are not so problematic as long as relative differences can be detected, this merely sidesteps the bigger problems of the validity and reliability of the measurement instrument. Furthermore, in cases where a single system is being evaluated such an argument does not hold.

10.4.1 Usability

To start, let us examine one of the most used conceptualizations of usability. The International Organization for Standards (ISO) [141, p. 2] defines usability as the extent "to which a product can

be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction". Thus, effectiveness, efficiency and satisfaction are identified as the key dimensions of usability. This definition also highlights the importance of carefully defining user and task models since it emphasizes the articulation of *specified* users and *specified* tasks. Nielsen [202] defines *usability* as "a quality attribute that assesses how easy user interfaces are to use,"[2] and divides the concept into five dimensions: learnability, efficiency, memorability, errors, and satisfaction. The ISO definition is arguably the most commonly used definition of usability and will be used in this paper. The ISO standard divides usability into effectiveness, efficiency and satisfaction. Definitions of these concepts are presented below.

- *Effectiveness* is the "accuracy and completeness with which users achieve specified goals." In other words, a tool is effective if it helps users accomplish particular tasks.
- *Efficiency* is the "resources expended in relation to the accuracy and completeness with which users achieve goals." A tool is efficient if it helps users complete their tasks with minimum waste, expense or effort.
- *Satisfaction* is the "freedom from discomfort, and positive attitudes of the user to the product". [141, p. 2]. Satisfaction can be understood as the fulfillment of a specified desire or goal. It is often the case that when people discuss satisfaction they speak of the contentment or gratification that users experience when they accomplish particular goals.

To understand the range of measurement techniques that researchers have used to study usability, Hornbæk [132] conducted a content analysis of 180 studies published in the core HCI journals and proceedings and classified the measures using the ISO definition of usability. Hornbæk [132] found that the way in which each of these concepts is operationalized varies according to study purpose. One important finding was that often multiple items are used to get at

[2] This definition is arguably not very good since *use* is the root of usability and it is used in the definition.

any one of these dimensions (this is common for all self-report data). For example, instead of a single item to measure satisfaction, several items are used to measure the construct in different ways. In general, measurement theory suggests the use of multiple items to measure a concept; single items are not considered reliable enough [105, 286]. The use of multiple items allows researchers to check for response reliability, as well as some types of validity.

10.4.1.1 Effectiveness

Hornbæk [132] found that the most common way that *effectiveness* has been measured in HCI studies is according to error rate and binary task completion. In IIR, the most common way to measure effectiveness is using traditional performance measures, such as precision and recall and to elicit self-report data from subjects about their perceptions of performance. In addition to error rate and binary completion, Hornbæk [132] documented several other measures of effectiveness including completeness, precision (ratio between correct information and total information retrieved), and recall (subjects' ability to recall information from the interface). While precision is operationalized the way it is in IR, recall is not. Also note that error rate, one of the most common measures, is not used that often in IIR except perhaps for fact-finding tasks, so there are some differences in notions of usability in HCI studies and usability in IIR evaluations. However, error rates with respect to interaction and interface use are worth capturing in IIR evaluations. Although many interfaces are quite simple, with more complex interfaces subjects may be observed making a number of errors, such as false clicks. Quality of outcome and expert assessment were also identified by Hornbæk; these are discussed in more detail later.

10.4.1.2 Efficiency

One of the most common ways to measure *efficiency* is to record the time it takes a subject to complete a task [132]. This includes measures of overall time, as well as more precise measures which document the amount of time subjects spend doing different things or in different modes. These types of measures were discussed previously

in Section 10.3 because like effectiveness, they are typically used in IIR evaluations as measures of performance. Usage patterns are also included in this set of measures, although in this discussion they have been separated out as interaction measures. In addition to these measures, questionnaire items regarding efficiency and time can be created to elicit subjects' perceptions.

Hornbæk [132] includes task difficulty as an efficiency measure where difficulty is typically determined by experts. While task difficulty is related to time-based measures of efficiency (i.e., presumably it takes longer to complete a difficult task), task difficulty was included in this paper as a contextual measure since IIR researchers are typically interested in seeing how difficulty impacts behaviors and performance. Moreover, task is a central focus in many IIR evaluations, so it is useful to separate it from efficiency since it is studied in a variety of ways in IIR. Learning measures are also identified by Hornbæk [132]. These measures use changes in efficiency as indicators of learning — e.g., subjects becoming faster at text input over time or subjects taking less time to complete subsequent tasks. These types of learning measures have not been used a lot in IIR evaluation.

10.4.1.3 Satisfaction

In the traditional conceptualization of usability, effectiveness, and efficiency measures are typically objective measures. The third dimension — satisfaction — attempts to gauge subjects' feelings about their interactions with the system. Hornbæk [132] identifies system preference as a measure of satisfaction, although in IIR evaluations system preference is typically treated as a separate construct and reported alongside satisfaction rather than as an indicator of satisfaction. Although multiple items are often used to assess satisfaction, a general question about satisfaction (e.g., how satisfied are you with your performance?) is also usually included as a questionnaire item. Specific satisfaction items might be asked for each different experimental feature of the system. Subjects' perceptions of outcomes and interactions are also commonly elicited. Examples include questions that ask about satisfaction with retrieval results and/or with the system's response time.

10.4.1.4 Ease of Use, Ease of Learning, and Usefulness

There are many other types of constructs that researchers try to measure in IIR evaluations that are related to usability. These include ease of use, ease of learning, and usefulness. Ease of use is considered by some as an indicator of satisfaction, but it can also be defined as the amount of effort which subjects expend executing and/or accomplishing particular tasks. Ease of use is closely related to efficiency: if a tool is not easy to use, then it is likely to result in inefficient use.

Ease of learning is related to the amount of effort subjects expend learning to use a system. Ease of learning items typically attempt to answer questions about how hard a system is to learn to use. A system may be effective and efficient, but if it is intended to have a human user and that user cannot learn to use it because it is too complex, then the system may not be successful.

Finally, usefulness is related to whether a tool is appropriate to the tasks and needs of the target users. The tool may be effective and efficient, but if users have no use for it, then it has little impact.

10.4.1.5 Available Instruments for Measuring Usability

There are several instruments that have been developed to measure usability. Each of these instruments has undergone different amounts of testing with regard to validity and reliability. Some of these instruments cost money while others do not. Some of these instruments are appropriate to IIR, but most contain a lot of items that do not make sense in the IIR evaluation context. Many of these questionnaires were developed in industry (and many are industry standards), but most questions are too general to provide enough detailed information about the IIR situation.

One of the most well-known instruments for measuring satisfaction is the Questionnaire for User Interface Satisfaction (QUIS) [53], which elicits evaluations of several aspects of the interface using a 10-point scale, including the subject's overall reactions to the software, the screen, the terminology and system information, and learning and system capabilities. The USE questionnaire [188] evaluates four dimensions of usability: usefulness, ease of use, ease of learning,

and satisfaction. Each dimension is assessed with a number of items which subjects respond to with a seven-point scale. Another commonly noted usability questionnaire is SUMI (Software Usability Measurement Inventory) [256]. SUMI consists of 50 items and provides subjects with three coarse responses: agree, do not know and disagree. Wildemuth et al. [295] used several validated usability measures in a TREC VID interactive project including Davis' [69] measures of perceived usefulness, ease of use and acceptance. Davis' work is from the management information systems (MIS) research, which has more examples of validated usability measures for information system evaluation.

10.4.2 Preference

In studies of two or more systems with a within-subjects design, it is common to collect preference information from subjects. This is perhaps one of the most basic types of data that can be collected, but often provides the clearest indication of subjects' attitudes about systems. Typically at least one open-ended follow-up question is asked to obtain more insight into subjects' preferences. Although not as common, preference data can also be elicited about individual aspects of the systems as well, such as method of display. Thomas and Hawking [270] propose an evaluation method that is based on preference. In this method, subjects are presented with a split screen each displaying search results from two different search engines. Subjects are asked to make holistic evaluations, basing their preferences on entire lists rather than individual documents.

10.4.3 Mental Effort and Cognitive Load

Mental effort is a construct that has been used as an indicator of efficiency, although the construct itself is quite complex [117]. This construct has been extensively investigated by those who study human computer interaction in the areas of control systems and airplane cockpits. Jex [151, p. 11] proposes the following definition of mental workload, "mental workload is the operator's evaluation of the attentional load margin (between their motivated capacity and the current task demands) while achieving adequate task performance

in a mission-relevant context". Hart and Staveland [122, p. 140], who developed the NASA-Task Load Index (NASA-TLX), define workload as "the cost incurred by a human operator to achieve a particular level of performance". Hart and Staveland [122, p. 140] go on to state that workload is "not an inherent property, but rather emerges from the interaction between the requirements of a task, the circumstances under which it is performed and the skills, behaviors and perceptions of the operator".

The NASA-TLX consists of six component scales, which are weighted to reflect their contribution to the workload according to the subject. These six scales are then averaged to produce an overall measure. The six factors that comprise the NASA-TLX are: mental demand, physical demand, temporal demand, performance, frustration and effort. A separate instrument is used for the weighting. This instrument elicits rankings of the *relative importance* of the six factors. This includes all possible pair-wise comparisons of the six factors. The number of times that a factor is rated as more important during the pair-wise comparisons indicates the relative importance of the factor. The NASA-TLX has been used in a few IIR evaluations. Most recently Kelly et al. [168] used this instrument to evaluate interactive question–answering systems and found that while it provided a good indication of mental effort for individual systems, it was hard to compare systems since the TLX is designed to distinguish among tasks, not systems.

In many traditional studies of mental workload a common method is to ask subjects to complete auxiliary tasks in addition to the primary task [151]. By manipulating the amount of cognitive load a person experiences with the auxiliary task and observing user performance, the belief is that it is possible to get an idea of what parts of the primary task are the most demanding, or alternatively, the most engaging. As the difficulty of the primary task increases or as the subject's engagement with the task increases, their performance on the auxiliary task should decrease because there are fewer cognitive resources available. Dennis et al. [72] employed the dual task technique in their study of interactive Web search. Dennis et al. [72] identified a number of possible auxiliary tasks and decided on a digit-monitoring task. The researchers found mixed results with respect to the usefulness of the

dual monitoring task, but the work is interesting in that it was one of the first uses of this technique in IIR. However, it is important to note that it is not always clear what differences in user behavior mean in these types of studies — increased effort on a task might mean the task is more difficult or that the user is more engaged. What can be said is that some tasks consume more attention that others and this can be good or bad.

10.4.4 Flow and Engagement

Although flow and engagement have not been used a lot in IIR research, they suggest additional ways that IIR systems and users' experiences can be evaluated. The notion of flow was proposed by Csikszentmihalyi [63] and is defined as a "mental state of operation in which a person is fully immersed in what he she is doing, characterized by a feeling of energized focus, full involvement, and success in the process of the activity." Csikszentmihalyi [63] identified five characteristics of the experience of flow and Csikszentmihalyi and Larson [64] presented an instrument and method for measuring flow in everyday life. Bederson [18] related flow to human–computer interaction and system evaluation. Pilke [213] conducted interviews to see if flow experiences occurred during IT use, while Chen et al. [52] looked at flow in the context of Web interactions. Ghani et al. [107] developed four seven-point semantic differential items to measure flow in human–computer interaction scenarios and further explored its relationship to task characteristics [106].

Engagement is a relatively new concept that has not yet been used to evaluate users' experiences with IIR systems. O'Brien and Toms [205, p. 949] define engagement as, "a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, and interest and affect". Thus, engagement is a multi-faceted construct that encompasses many characteristics of a user's experience. O'Brien and Toms describe how the theory of engagement shares some attributes with flow, aesthetic, play and information interaction theories, but state that it is fundamentally different from these theories. O'Brien and Toms [205] provide a conceptual framework

for defining user engagement and have also developed a procedure for measuring user engagement [206].

10.4.5 Subjective Duration Assessment

As described earlier, subjects often exhibit a number of response biases when responding to self-report measures which can make it difficult to obtain valid and reliable data. Motivated by the finding that subjective and objective performance measures are often uncorrelated and a belief that this indicates a potential problem with the validity of self-report measures, Czerwinski et al. [65] explore the use of time estimation as a way to obtain an indirect measure of task difficulty from subjects. Czerwinski et al. [65] call this method of estimation *subjective duration assessment* and propose a measure called *relative subjective duration*. After completing tasks in Czerwinski et al.'s [65] study, subjects were asked to estimate the length of time it took them to complete tasks. This estimation was then compared to the actual length of time it took them. Czerwinski et al. [65] found that subjects underestimated times associated with tasks with high success rates and overestimated times associated with tasks with low success rates. Relative subjective duration provides an alternative method to elicit estimations of subjects' perceptions of task difficulty. While this only represents one such measure, it suggests a possible useful approach to discovering other indirect ways of assessing subjects' experiences with IIR systems.

10.4.6 Learning and Cognitive Transformation

As described earlier, IIR is typically not a goal unto itself, but is often done in support of some larger goal or task. One goal of most IIR systems that is implied rather than explicitly stated is that the system will help users learn about a particular topic. This, of course, is accomplished through retrieving, reading and evaluating documents. However, measuring the extent to which learning has taken place is difficult because it would require the establishment of some baseline measure of how much a subject knows about a topic and a post-test measure to determine how much they have learned. One example study that attempted to do this is Hersh et al. [130]. Developing some standardized

instrument to evaluate how much a person knows is not easy and each different topic might require a different assessment technique.

Collecting assessments of interaction outcomes (e.g., a paper) is another approach to gauging the extent to which a system helped a person learn about a topic or accomplish a goal. This is not technically a self-report measure, but such assessments are usually generated by experts or other subjects who use standardized instruments to make assessments. As noted earlier, example studies that have attempted to assess final products include Egan et al. [83], Kelly et al. [166], Marchionini and Crane [192] and Vakkari [280].

11

Data Analysis

This section focuses primarily on quantitative data analysis, since much of the data collected in IIR evaluations is quantitative in nature. Some of the data collection techniques described above yield qualitative data, so qualitative data analysis will be presented briefly. This brevity is not meant to indicate that qualitative data analysis is easier or less important, but rather that there just is not enough space to provide a detailed exploration of this topic.

11.1 Qualitative Data Analysis

While there are numerous approaches that one might take in doing qualitative research — many of which differ epistemologically and philosophically (see, for instance, [142]) — this article focuses on two of the more common approaches to qualitative data analysis, content analysis and open coding. Since interviews only form a small part of traditional IIR evaluations, the two techniques described below should provide adequate background for analyzing this type of data. Those interested in reading more about different qualitative research traditions and analysis techniques are referred to Miles and Huberman [195],

Denzin and Lincoln [73], Charmaz [50] and Glaser and Strauss [108]. References from the information science literature include Bradley [40], Dervin [74] and Fidel [91].

The goal of most qualitative data analyses that are conducted in IIR is to reduce the qualitative responses into a set of categories or themes that can be used to characterize and summarize responses. Perhaps the most important message that can be communicated about qualitative data analysis is that it is not as easy as it seems. In general, reports of qualitative data analyses in IIR are weak and usually inadequate. One reason for this is that those unfamiliar with qualitative data analysis often do not bother to report important details about how the data were collected and analyzed. For instance, consider the following three scenarios:

(1) A researcher records and transcribes an interview. Analysis is based on the transcriptions.
(2) A researcher records an interview, but does not transcribe it. Instead the researcher listens to the recordings once and takes notes. Analysis is based on these notes.
(3) A researcher does not record the interview, but takes notes during the interview. Analysis is based on these notes.

All things being equal, the quality of the data the researcher captures as well as the researcher's interpretations of this data in each of these scenarios is likely to vary. In Scenario 1, the recording and transcription processes will result in the most faithful record of what occurred during the interview. Thus, analysis based on this data will likely be better than if these steps were not taken. In Scenario 2, an accurate recording of what occurred during the interview exists, but there is no guarantee that the researcher has done a good job noting what occurred, which of course, has implications for the validity of the subsequent analysis. In Scenario 3, the representation of what occurred will be limited to what the researcher can physically record and also to what the researcher thinks is important to record at the time of the interview. The selectiveness of this process will be reflected in the analysis. It is also the case that one's ability to conduct a good interview will be compromised

since one is engaged in both interviewing and note taking. The point of this example is not to say that all interviews should be recorded and transcribed, but rather to point out the potential differences of each method and illustrate the importance of reporting the method in its entirety even if it *seems* trivial.

Another reason that reports of qualitative data analysis are typically weak in IIR is because researchers often do not make appropriate distinctions among different analysis techniques. For instance, it is common for researchers to use the term *content analysis* as a synonym for qualitative data analysis. Researchers use this word in a generic sense — there is content and it needs to be analyzed — but content analysis actually represents a very specific data analysis technique. While content analysis and qualitative analysis have some things in common, they represent two very different approaches to analyzing textual data.

Content analysis is most often used to analyze recorded communication — books, films, email messages, Web pages, and advertisements. At its inception, it was intended as a quantitative method, although there are now a number of variations, interpretations and uses of content analysis [201]. Content analysis was originally executed in much the same way that IR is executed — by counting the occurrences of words and other features. Traditional content analysis starts with a somewhat well-defined and structured classification scheme, including categories and classification rules. The categories are usually mapped to variables. For example, if one were analyzing a set of transcripts for mentions of the concept of relevance, then one might use a pre-defined vocabulary as indicators of this concept. Before the analysis can start, the researcher creates a codebook that links together the concepts of interest, the categories that represent them and the classification rules. The coding process is more structured and deductive than what it is in qualitative data analysis.

Most researchers in IIR engage in a less structured form of data analysis when analyzing qualitative data. The goal is still data reduction, but the process differs from content analysis in several keys ways. The codes and categories are usually developed inductively during the analysis process as the researcher analyzes the data. This process is

referred to as *open-coding* [193]. Strauss and Corbin [253, p. 62] characterize open coding as, "the part of analysis that pertains specifically to the naming and categorizing of phenomena through close examination of data ... during open-coding the data are broken down into discrete parts, closely examined, and compared for similarities and differences". Codes are suggested by the researcher's examination and questioning of the data. This process is iterative; when new codes are added previously categorized data are reviewed to see whether they need to be reclassified. Coding ceases when saturation has been reached and all relevant utterances have been classified. While researchers typically develop rough heuristics for classifying data into different categories, these are not as well-formed as those in content analysis, which has implications for reliability. With content analysis, some type of inter-coder reliability should be performed to ensure that items have been coded consistently. With open-coding, this step is not always required, expected or possible, although it is assumed that the researcher is analyzing the data consistently and faithfully.

The two approaches discussed in this section are not the only approaches to analyzing qualitative data. In some ways they represent two ends of a continuum. On one end is content analysis, which is highly structured and emphasizes reliability, and on the other end is open-coding, which is more fluid and emphasizes flexibility. Both of these approaches, as well as all the ones in between, can be used to analyze qualitative data in IIR evaluations.

11.2 Quantitative Data Analysis

Quantitative data analysis is a large and complex topic. In this section, basic statistical tests are presented which are used in the common IIR evaluation model where a researcher is comparing two or more systems or interfaces (independent variable) using a set of outcome measures (dependent variables) that are categorical or continuous in nature. Reading this section will not make anyone an expert, but it will help readers distinguish among different types of statistics, select appropriate statistical tests and understand how some statistics are computed. The following books were consulted during the writing of

this material: Cohen [56], Gravetter and Wallnau [110], Myers and Well [200] and Williams [296].

A statistic is an estimate of an unknown value in the population; these unknown values are known as parameters. Statistics are derived from samples and provide *estimates* of the values of unknown parameters in the population. Descriptive statistics characterize variables; most notably these statistics describe central tendency and variation. Descriptive statistics are the basic inputs of inferential statistics. Inferential statistics are used to compare the relationship among two or more variables and to test hypotheses. Inferential statistics allow one to make *inferences* about population parameters based on sample statistics.

Inferential statistical tests are often performed in order to determine whether null hypotheses can be rejected and *significant* (or *statistically reliable*) relationships exist among variables. The word *significant* in the context of statistics has a specific meaning; *significant* is used to describe situations where particular probability values are observed. Thus, *significant* should not be used as a synonym for *large* or *important* when presenting and discussing results. If it is claimed in a research report that a *significant* relationship was observed, then one should be prepared to present the statistical tests supporting this claim.

The inferential statistics reported in this paper are commonly used in IIR evaluations. In particular, the focus is on tests that are used in evaluations to compare two or more systems among a set of outcome measures. The statistics reported in this section also focus on parametric statistics rather than non-parametric statistics. Parametric statistical tests assume that the variables that are being examined are normally distributed in the population (this will be discussed in more detail later). When it is assumed that the variable is not distributed normally in the population, then non-parametric tests can be used. Non-parametric tests use different descriptive statistics in their computation than their parametric counterparts. For instance, two medians might be compared instead of two means. When appropriate, an equivalent non-parametric test is suggested for each parametric test.

The research scenario in Figure 11.1 is provided to facilitate the presentation of material in this section. This scenario is modeled after

A researcher has developed two experimental IR systems and would like to test them against one another and a baseline. These three systems will be called System A (the baseline), System B and System C (note that *system type* functions as one variable, with three levels). Subjects are given six search tasks to complete which ask them to find documents that are relevant to pre-determined topics. Each subject completes two searches on each system (a within subjects design).

The researcher is interested in comparing the three systems using the measures listed below. These measures are organized according to the instrument used to collect the data.

Demographic Questionnaire

Sex of Subject [Male or Female]

Pre-Task Questionnaire

How *familiar* are you with this topic? [5-point scale, where 1= know nothing about the topic and 5=know details]

Post-System Questionnaire

Usability [5-point scale, where 1=strongly disagree and 5=strongly agree]: It was easy to find relevant documents with the system.

Exit questionnaire

Preference: Which system did you prefer? [System A, B or C]

System Logs

Performance

- Average session-based nDCG

Interaction

- Number of queries issued
- Query length

Fig. 11.1 IIR Research Scenario.

the archetypical IIR evaluation and compares three systems.[1] However, it should not be used as a self-contained model since it only represents sample variables and is necessarily incomplete. The different variables have been purposely selected to illustrate the statistics that are discussed.

11.2.1 Descriptive Statistics

The first step in analyzing data is to examine the frequency distributions of each variable. This is useful for identifying outliers and anomalies, and human errors that may have been made during the process of building the data files. It is also useful because it helps one understand the appropriateness of different types of statistics since this depends in

[1] *System* is used in a generic sense. A researcher may also be comparing two interfaces or interaction techniques.

part on the distributions. Frequency distributions present the number of observations of each possible value for a variable. For example, the frequency distribution for the familiarity variable will show how many times each of the five-points was used by subjects.

Six of the most common types of distributions are shown in Figure 11.2. These are the (a) normal distribution (also known as the bell-shaped curve and Gaussian distribution); (b) peaked distribution; (c) flat distribution; (d) negative skew; (e) positive skew; and (f) bimodal distribution. These curves represent general, theoretical shapes. The normal, peaked, flat and bimodal distributions are symmetrical and the negative and positive skews asymmetrical. While the distributions of some real data will match these shapes nearly perfectly, most distributions only approximate them.

There are two main measures for describing a curve's shape: skew and kurtosis. For each measure, a value of zero represents a normal curve. The skew measures can be best illustrated with the negative and positive skewed distributions. The skew measure will move in a negative direction when the distribution approximates a negative skew and positive direction when it approximates a positive skew. Kurtosis can be illustrated with the peaked and flat distributions. The kurtosis measure will move in a negative direction when the distribution approximates

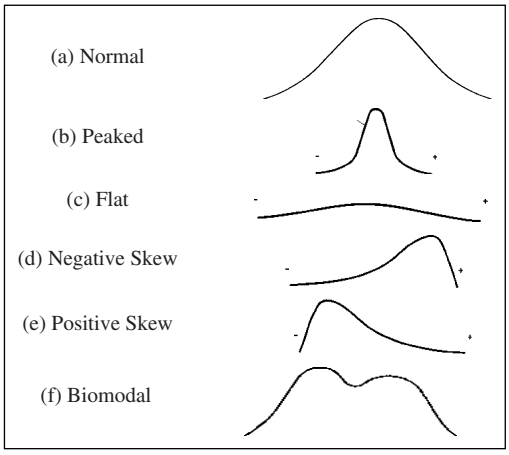

Fig. 11.2 Example distributions.

the flat curve (platykurtic) and a positive direction when it approaches the peaked curve (leptokurtic). A normal curve, which has a kurtosis of zero is said to be mesokurtic. These measures can be particularly useful when one is transforming or re-classifying data. For instance, if one decides to transform data collected using a seven-point measure into a three-point measure and is interested in having roughly equal numbers of observations for each point, then the kurtosis measure can be used to identify which of the candidate divisions of the seven-point measure results in the flattest distribution. Variables that are skewed can also be subjected to mathematical transformation in order to decrease the differences between points. For instance, a logarithm transformation can be performed (if appropriate) to minimize the differences between scores. We will discuss these distributions in more detail below, but first let us review some other descriptive statistics.

11.2.1.1 Measures of Central Tendency

There are two major classes of descriptive statistics: measures of central tendency and measures of dispersion. These statistics are useful for describing distributions so that the entire frequency curve does not have to be presented. Measures of central tendency describe how observations (or scores) cluster around the center of the distribution. There are three basic measures: mean, median, and mode. The definitions of each are provided below (Table 11.1). To illustrate these measures, let us assume that we have data from five subjects, each of whom submitted the following number of queries: 2, 1, 3, 5, and 1. The central tendency measures for this sample are: Mean = 2.4, Median = 2 and Mode = 1.

For the normal curve these measures are equal, but in distributions of different shapes the measures are not equal and can give misleading

Table 11.1 Measures of central tendency: mean, median and mode.

Central Tendency	
Mean	The sum of the scores in a distribution divided by the total number of scores. $\overline{X} = \frac{\sum X}{n}$
Median	The score that falls in the middle of the distribution.
Mode	The score that occurs the most frequently.

representations if used to describe the data. For a distribution that is positively skewed, the mean will represent a value that is higher than most other scores, especially the most common ones. The mean will also be greater than the median. For a distribution that is negatively skewed, the mean will represent a value that is lower than most other scores and will be slightly less than the median. In a bimodal distribution, there are two modes. The mean and median will be near the ebb where the two modes (or humps) come together. The mean, median, and mode are approximately equal in peaked and flat distributions.

Some measures are more or less appropriate for describing different variables given the level of measurement. For instance, it does not make sense to report the mean sex of subjects since sex is a nominal variable. Even though the researcher will probably assign numeric values to the two levels of sex to facilitate analysis, these serve no other function than assisting with computing frequencies. Thus, for nominal variables, the most appropriate measure of central tendency is the mode. One should also exercise caution when reporting the mean of an ordinal level variable when it represents categories.

11.2.1.2 Measures of Dispersion

While measures of central tendency describe where scores cluster in a distribution, measures of dispersion describe how scattered scores are about the center or how scores deviate from the mean. There are three basic measures of dispersion: range, variance, and standard deviation. Definitions of each are provided below (Table 11.2). Variance and the standard deviation are very similar. The major difference is that the standard deviation is smaller (since it is the square root of the variance). The dispersion measures for our sample data are: range = 4, variance = 2.80, and standard deviation = 1.67.

An earlier distinction was made between formulas for samples and for populations. In the formula for computing variance, if a population was studied instead of a sample, the denominator would be N instead of $n - 1$ (note that the capital N is used to indicate the size of the population, while the lower case n is used to indicate the size of the sample). There are different ways to note values that describe

Table 11.2 Measures of dispersion: range, variance, and standard deviation.

Measures of Dispersion		
Range	The difference between the maximum and minimum scores in a distribution.	$\text{Max}_x - \text{Min}_x$
Variance	The mean of the squared deviation scores. To compute the variance, first compute the mean for a set of numbers and then subtract each individual score from this mean. Square these values and then sum. This value is called *sum of squares*. Next, divide the sum of squares by the total number of scores minus 1 $(n - 1)$.	$\frac{\sum (\text{X} - \overline{\text{X}})^2}{n-1}$
Standard Deviation	The square root of the variance.	$\sqrt{\frac{\sum (\text{X} - \overline{\text{X}})^2}{n-1}}$

Table 11.3 Notations for descriptive measures of populations and samples.

	Population	Sample
Mean	μ	$\overline{X}$
Variance	σ^2	s^2
Standard deviation	σ	s
Number of observations	N	n

populations (parameters) and values that describe samples (statistics). This notation is provided in Table 11.3.

Subtracting 1 from the sample size is common in many formulas. It can be thought of as a penalty to ensure that the sample values are unbiased estimators of the population values. This is related to *degrees of freedom. Degrees of freedom* is the extent to which scores in a sample are free to vary. In our example set of query scores, the mean number of queries is 2.4 and the sum of all the values is 12. Degrees of freedom do not change this summative value (whatever it might be), but say something about how much freedom individual scores have to vary and still yield the same sum. That is, in order to arrive at the same sum of scores, each of the individual scores can vary so long as the value of one score is fixed. Any number of scores can be added and yield the value of 12, but one score will have to be reserved to maintain this. For instance, the 1st–4th query scores could be 2, 4, 4, and 1, which would add to 11. This is would mean that in order to maintain a sum of 12, the 5th score would have to be 1. Thus, given a particular sum of values, the values of all of the scores except one can vary. Degrees

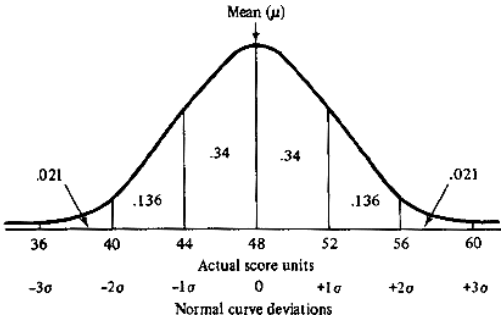

Fig. 11.3 The normal curve and its properties with respect to distribution of data.

of freedom indicate how many scores can vary and how many have to stay fixed.

The curve shown in Figure 11.3 is useful for illustrating dispersion and particularly, the standard deviation. The normal curve has special properties that allow one to make statements about how scores fall around the mean with respect to different units of the standard deviation. These properties are displayed on the normal curve in Figure 11.3. In normal distributions, approximately 68% of all scores fall within $+/-1$ standard deviation of the mean, 95% fall within $+/-2$ standard deviations and 99% fall within $+/-3$ standard deviations of the mean. For any item that is normally distributed in the population, these characteristics will hold. Thus, knowledge that the variable of interest is distributed normally within the population and knowledge of the mean and standard deviation of the sample allows one to understand the distribution of all scores.

Let us consider an example. Imagine if a researcher found that the mean score for the usability measure listed above was 3.5. This value provides an indication of the average value subjects marked for this question. Now imagine that the standard deviation was 0.25. This would indicate that the scores were fairly similar to one another; that is, that about 68% of the scores were between 3.25 and 3.75. Now imagine that the standard deviation was 2.0; this value would indicate that scores were fairly different from one another and that subjects used a range of values. The same number of scores (68%) would fall between the values of 1.5 and 5.5. For variables that are normally distributed

in the population, the mean and standard deviation allows one to form a basic understanding of the data. Thus, when reporting measures of central tendency some measure of dispersion should be reported. Each statistic provides an important and different characterization of the data. Reporting a measure of central tendency such as the mean without reporting a measure of dispersion can give misleading results.

Measures of dispersion describe the distribution of scores about the central tendency. If the standard deviation is large relative to the mean, then scores will be more varied. If the standard deviation is small relative to the mean, then scores will be more alike. This can be easily illustrated with the peaked and flat distribution in Figure 11.3. In a peaked distribution, scores are stacked and tightly clustered around the mean and as a result the standard deviation is small in relation to the mean. In a flat distribution, scores do not cluster and are spread evenly across a range of values; as a result, the standard deviation is large in relation to the mean.

One of the most common concerns that researchers have when preparing to conduct data analysis is that the distributions of *their* data are not normal. This is not unusual, since sample sizes are typically quite small and probability sampling techniques are not used. Since researchers are usually working from small samples, it is unlikely that the variable *in their sample* will be normally distributed. Consider the following illustration of this point. Two-hundred researchers across the world decide to collect data about the same variable in the same way. Each researcher can create an individual distribution curve from the data they have collected (a sample distribution). Each of these distributions is likely to vary, but when they are put together they should form a normal distribution (if the variable is in fact normally distributed in the population). By themselves, each of the 200 sample distributions may look strange, but many will be very similar — these distributions will form the hump of the normal curve. Many will be unique and differ from most others — these distributions will form the tails of the normal curve. Thus, depending on the sample distribution, one's sample data may or may not reflect a normal distribution. However, the requirement is that the data is distributed normally in the population.

11.2.1.3 *z*-scores

A common way to standardize and compare scores across a range of distributions is to create *z-scores*. Raw scores are transformed relative to the mean of the scores, so that the mean is set to the value zero on the normal curve in Figure 11.3. A z-score describes the exact location of a score within a distribution: the sign of the score ($-$ or $+$) indicates if the score is above or below the mean and the magnitude of the score indicates how much the score is above or below the mean. The formula for transforming raw scores into z-scores is

$$z = \frac{X - \mu}{\sigma}$$

where X is the raw score, μ is the mean of the all of the scores, and σ is the standard deviation. It is important to point out that in this formula population parameters are used for the mean and standard deviation (notice the use of μ and σ). The value produced by the formula situates the individual score relative to the mean and states how many standard deviations the raw score is from the mean and in which direction. The numerator in the equation is known as the *deviation score*; these values are used in a number of inferential statistical tests and will be re-visited.

Because z-scores rest on an assumption of normality, information about the likelihood of any particular score occurring in a population is available. Specifically, it is possible to indicate the probability of obtaining a score that is higher or lower than a particular target score. As described earlier, the normal distribution is symmetrical and has certain properties with respect to how scores are distributed about the mean. The most basic statement that can be made is that the chance of observing a score above the mean is 50% and the chance of observing a score below is mean is also 50%. Note that the z-score for a mean is equal to zero. The chance of observing a raw score that is greater than a z-score of $+1$ is about 16/100 since only about 16% of all scores are in this part of the curve. The chance of observing a raw score below this z-score is about 84/100. Thus, z-scores can be used to reason about the likelihood of observing particular scores by turning percentages into proportions. In this example, the z-score is an integer which can be

used easily with the normal distribution to find probability values for observing particular scores. In cases where the z-score is not an integer, the *unit normal table* can be used [27]. Portions of this table are shown below in Table 11.4.

Reasoning about the likelihood or probability of some event occurring is the basis of statistical inference. In the table above, the probability of observing a z-score of 1.65 is less than 0.05. The reader may be familiar with seeing probability values of 0.05 and 0.01 used in conjunction with statistical tests. These values refer to particular areas of the normal curve and the likelihood that some observed sample mean will fall in this area. The z-score formula in the preceding paragraph assumes that the researcher has knowledge of two population parameters — the mean and standard deviation. Of course, this is rarely the case in any study, especially IIR evaluations. In the next section, we will examine several examples that allow us to move from knowing some information about the population to not knowing any. This will allow us to look closely at some important statistical concepts, in particular those that try to account for discrepancies between sample statistics and population parameters.

Sampling error is the amount of error between a sample statistic and its corresponding population parameter. Thus, when computing inferential statistics one must somehow account for the error introduced by studying a sample instead of the entire population. As described previously, it is not always the case that a researcher's sample data will be distributed normally. However, an assumption is made that if many researchers collected many different samples from the same population, then the distribution of means for all the samples together will be normally distributed. That is, most researchers will find similar means; these will pile-up and form the hump on the normal curve. Some researchers will find means that vary different distances from the common mean; these values form the tails of the curve.

These ideas form the basis of the *central limit theorem*, which is at the core of most inferential statistical tests. This theorem states that the distribution of sample means will approach a normal distribution displaying the real mean and standard deviation of the population as

Table 11.4 Unit normal table for interpreting z-scores.

z-score	Proportion in body	Proportion in tail		z-score	Proportion in body	Proportion in tail		z-score	Proportion in body	Proportion in tail
0.00	0.5000	0.5000		1.00	0.8413	0.1587		1.64	0.9495	0.0505
0.01	0.5040	0.4960		1.01	0.8438	0.1562		1.65	0.9505	0.0495
0.02	0.5080	0.4920	...	1.02	0.8461	0.1539	...	1.66	0.9515	0.0485
0.03	0.5120	0.4880		1.03	0.8485	0.1515		1.67	0.9525	0.0475
0.04	0.5260	0.4840		1.04	0.8508	0.1492		1.68	0.9535	0.0465
0.05	0.5199	0.4801		1.05	0.8531	0.1469		1.69	0.9545	0.0455

n (the number of samples) approaches infinity. The basic idea behind this theorem is that as n gets larger, the distribution of sample means more closely approximates the normal curve — as a result the error between the sample and population means decreases. In the section on sampling, the relationship between sample size and power was discussed. This is embodied in the *law of large numbers*: larger samples will more representative of the population from which they are selected than smaller samples.

11.2.2 Inferential Statistics

As mentioned above, inferential statistics allow one to make inferences about population parameters based on sample statistics. Inferential statistics are most often used to test relationships between two or more variables and evaluate hypotheses. While it is possible to test the difference between a known population parameter and a single variable, it is rarely the case that the population parameter is known. Although uncommon, researchers who perform inferential statistics are meant to select appropriate tests during the design phase of a study. Thus, inferential statistics can be viewed as research tools about which researchers make choices, just as other instruments. Furthermore, the choice of which test to use is determined in part by variable data types (levels of measurement), so considering these things simultaneously will likely lead to better design choices.

11.2.2.1 *z*-statistic

Similarly to how one uses z-scores to reason about individual scores, one can also use z-scores to reason about the likelihood of observing particular sample means. The z-score formula changes slightly since we are dealing with sample means instead of individual raw scores. The z-score formula for reasoning about the likelihood of sample means is

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \quad \text{where } \sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}}.$$

The numerator in this formula is virtually the same: it is the difference between the real and observed means. The denominator is a bit different and is known as the *standard error*. Instead of being equal to the standard deviation of the population, it is defined as the square-root of the population standard deviation squared divided by the sample size. Basically, the standard error is a measure of the difference that would be expected to occur by chance between the sample mean and the population mean. This value is clearly a function of the size of sample size: when the sample size increases the standard error gets smaller and the sample mean tends to more closely approximate the population mean.

Although the z-score formula of sample means still assumes that the researcher has knowledge of the population mean and standard deviation, it can be used in special cases for hypotheses testing when this information is available. There are not any apparent situations in IIR where this might be possible, so an example from education will be provided. The point of providing such an example is to segue from z-scores to test statistics. In fact, in this situation it is more appropriate to refer to the z-score as a *z-statistic* (similar to the t- and F-statistics for the t-test and ANOVA, respectively), since it is used for hypotheses testing using sample data.

Suppose a researcher has created a two-week program that is meant to educate a person about information literacy (IL). The researcher hypothesizes that this program will have an effect on IL. The null hypothesis is that the program has no effect, or stated another way that there is no relationship between the program and IL. Information literacy (IL) will be measured using a nationally standardized test. It is known from national results that scores on this test are normally distributed and that the average IL score is 70 ($\mu = 70$) with a standard deviation of nine ($\sigma = 9$). Suppose the researcher enrolls 16 people in the study ($n = 16$). After the program these people complete the IL test and score a mean of 74.

To evaluate whether the difference between the average IL score of study subjects is statistically different from the average IL score of the population, a z-statistic can be computed to determine the probability of the sample IL score occurring in the population. Before this statistic is computed, the researcher must determine the boundary that

separates the high-probability samples from the low-probability samples. This boundary is known as the *alpha level* and by convention is typically set to 0.05, although in medical research this value is often smaller (0.01 or even 0.001). This value should be recognizable to many readers as the probability value (p-value) that is typically reported alongside statistical results. This value will separate the most unlikely of the sample means (from our imaginary distribution of means) from the more common ones (95% of them).

The extremely unlikely values defined by the alpha level make up what is called the *critical region* of the curve — basically this area is at the extreme ends of the normal curve. Because the researcher did not state a directional hypothesis, but just stated that there would be an effect, the 0.05 must be split evenly so that half of the critical region (0.025) is at the low-end of the curve and the other half is at the high-end of the curve. This is known as a *two-tailed test* — the experimental educational program can have either a positive or negative effect on IL scores; the sample mean may be found in the extreme left (low) or right (high) tails of the curve. As stated earlier, the researcher's hypothesis was non-directional — that is, it did not state whether this program would have a positive or negative effect, so a two-tailed test is most appropriate. The alternative is a *one-tailed* or *directional hypothesis test* where the researcher states that the impact of the experimental treatment (i.e., the education program) will have a positive (or negative) effect on IL scores. For one-tailed tests, the critical region is contained within one area of the curve, either the extreme right (for positive effects) or the extreme left (for negative effects), and the 0.05 is concentrated in one end or the other (it is not split) thus making the critical region larger.

Essentially, the value that is being tested in a statistical test is the *difference* between the population mean and the sample mean (in other statistical tests it is the difference between two or more sample means). Assuming a constant alpha level, it is easier to achieve statistical significance with a one-tailed test, since the larger, concentrated critical region accommodates a larger number of values. The necessary minimum difference between means to make it into the critical region will be smaller, and thus, more easily achievable, with a one-tailed test

than with a two-tailed test (although one could adjust the alpha level for one-tailed tests). In many IIR evaluations, researchers are able to make directional hypotheses. However, the standard practice is to use two-tailed tests[2] with an alpha level of 0.05, which is basically equivalent to a one-tailed directional test with an alpha level of 0.025. This is not particular to IIR; in most behavioral sciences two-tailed tests are normally used even when directional hypotheses are stated because the risk of rejecting a null hypothesis that is actually true[3] is too great with one-tailed tests. Since a two-tailed test requires stronger evidence to reject a null hypothesis it provides more convincing results that the null should be rejected. All tests discussed in this paper will assume two-tailed tests.

Let us get back to the example of the educational researcher. First, we would consult the unit normal table to determine the critical region of the curve for $p = 0.05$. The critical value separating the uncritical region from the critical region of the curve is displayed in the portion of the unit normal table in Table 11.4. We see that z-scores that are greater than or equal to 1.65 are in the portion of the tail that corresponds to an alpha value of 0.05. The z-statistic for our example data is 1.77, so we can reject the null hypothesis since the probability that a z-score of 1.77 came from the population distribution is less than 5/100.

When conducting hypotheses tests, there are two important types of errors that can occur, *Type I* and *Type II*. A Type I error occurs when a researcher erroneously rejects the null hypothesis. Type I errors can occur because of the researcher's actions — e.g., blatantly ignoring test results or setting the alpha value too low — but the more common source of Type I errors are anomalous results. In these cases, results are found to be statistically significant with a standard alpha value of 0.05, but there is something peculiar about the sample or testing method that caused the results. In the IL example above, it may have been that all of the members of the sample were just smarter than average, which would explain why they scored higher. The alpha level

[2] This is the default in most statistics packages.
[3] This is known as a Type I error.

actually determines the probability that a statistical test will lead to a Type I error. For instance, an alpha level of 0.05 means that there is a 5% chance that the data obtained in the study was a result of some anomalous condition. In other words, if the study were done 100 times, the same results are expected 95 of those times.

While the risk of a Type I error is actually quite small for single hypothesis tests, when researchers conduct multiple independent hypotheses tests on the same data set, the critical alpha value is often adjusted to further safeguard against obtaining statistical significance by chance. One common type of alpha adjustment is the Bonferroni correction, which reduces the critical alpha value by dividing some standard, such as 0.05, by the number of independent hypotheses that will be evaluated. For example, if a researcher examines five independent hypotheses, then a Bonferroni correction would change the critical alpha value to 0.01 (0.05/5). A less restrictive correction is the Holm–Bonferroni method.

A Type II error is failing to reject the null hypothesis when it should be rejected. In these situations, the test statistic does not fall into the critical region of the curve, even though the treatment may have a small effect. Strictly speaking, there is no way to determine the probability of making such an error. Probability values can provide a hint that a Type II error has occurred (for instance, if the test statistic falls into a region defined by 0.07 instead of 0.05), but this is not a universal explanation for why significant results were not found. In some cases, researchers often claim that results are *almost* statistically significant, but this inappropriate. If, for instance, two more subjects were included in the sample, the test statistic might move *further away* from the critical region instead of closer to it. Repetition of an experiment will allow one to *explore* the possibility that a Type II error occurred, but it is not a guarantee that results will change.

11.2.2.2 *t*-statistic and t-tests

The z-statistic in the preceding example is useful in cases where the population mean and standard deviation are known *a priori*, but this

is rarely the case in IIR evaluations. It is usually the case in most IIR evaluations, that two or more sample means are being compared with no knowledge of the population parameters. For instance, in the IIR research scenario, a researcher might be interested in comparing the differences in performance according to sex — is there a difference between the performance of males and females? The formula used to compute the z-statistic can be modified to account for the differences in what is known (or unknown) about the population parameters. First, the denominator in the formula for the z-statistic changes to one based on the variance of the sample, instead of one based on the standard deviation of the sample. The new denominator is called the *estimated standard error*:

$$z = \frac{X - \mu}{s_{\overline{X}}} \quad \text{where } s_{\overline{X}} = \sqrt{\frac{s^2}{n}}.$$

The second change accommodates the comparison of two sample means, which basically doubles all of the elements of the formula above, so that the above formula becomes:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}.$$

In this formula, the two sample means are represented by $\left(\overline{X}_1\right)$ and $\left(\overline{X}_2\right)$. The null hypotheses, which assumes that these two samples were drawn from the same population and that there would be no difference between the means, is represented by the difference between μ_1 and μ_2 (which is eventually replaced by zero and excluded from the formula).

The magnitude of the standard error (the denominator in the formula) is determined by the variance of the observed scores and the sample size. With two sample means, there are two measures of variance (one for each sample) and therefore, two standard errors. In the formula above, the denominator is defined as:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}.$$

Research Question: What is the relationship between sex and perceptions of system usability?
Hypothesis: Males will rate the system as more usable than females.

	Sex		Deviations from samples means		Deviations squared	
	M	F	d_M	d_F	$d_M{}^2$	$d_F{}^2$
Usability scores	4	3	–0.2	1	0.04	1
	5	3	0.8	1	0.64	1
	4	2	–0.2	0	0.04	0
	5	1	0.8	–1	0.64	1
	3	1	–1.2	–1	1.44	1
Mean (M)	4.2	2				
n	5	5		Σd^2	2.8	4

$$t = \frac{\overline{X}_M - \overline{X}_F}{\sqrt{\left(\frac{\Sigma d_M{}^2 + \Sigma d_F{}^2}{n_M - n_F - 2}\right)\left(\frac{n_M + n_F}{n_M \cdot n_F}\right)}}$$

$$= \frac{2.2}{\sqrt{\left(\frac{2.8+4}{5+5-2}\right)\left(\frac{5+5}{5 \cdot 5}\right)}}$$

$$t = \frac{2.2}{.58309} = 3.77$$

Fig. 11.4 Calculation of *t*-statistic using sample sex and usability data.

This formula is based on the pooled variance of the two samples,[4] where

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}.$$

SS represents the *sum of squared deviations*, which is best exemplified in Figure 11.4. Recall that the sum of squares is used in the computation of the variance and standard deviation. Within each sample group, each observed score is subtracted from the mean and squared. These values are then summed. The other figures in the formula for pooled variance as well as the parent formula for the standard error are based on the sample sizes of the two groups being compared. Sample size figures into the calculation directly (i.e., using the number itself) and indirectly through degrees of freedom (*df*). As a reminder, degrees of freedom is the extent to which scores in a sample are free to vary. For the *t*-statistic, this is equal to, $n_1 - 1 + n_2 - 1$, where n_1 is equal to the number of observations in the first group and n_2 is equal to the number of observations in the second group. This is often abbreviated as $n - 2$, where n is equal to the total number of observations in the study (i.e, total sample size).

[4] This is basically an average of the variances for the two samples.

To consolidate the formulas above, we can replace the $s_{\overline{X}_1-\overline{X}_2}$ in the t-statistic formula with the actual formula:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}.$$

Figure 11.4 walks the reader through a complete example based on the IIR Research Scenario. In this example, the researcher is interested in investigating the effect of sex on usability. The t-statistic formula given in the example is a variation of the one above. Note that it does not include $(\mu_1 - \mu_2)$ since this value is assumed to be zero.

To determine whether a t-statistic is significant, the t-distribution table has to be consulted. This is similar to consulting the unit normal table that was consulted to determine if a particular z-statistic was significant. The objective is the same: to determine what t-value needs to be observed in order for the statistic to fall in the critical region of the curve.[5] To determine this value, we need to use the total degrees of freedom $(n - 2)$ and the alpha level. The example in Figure 11.4, $df = 8$ and the alpha level is 0.05. A portion of the t-distribution table [27] is presented in Table 11.5. To use this table, we first find the column that corresponds to an alpha level of 0.05 for a two-tailed test and then the row that corresponds to our degrees of freedom. The figure at this intersection tells us that we need a value that is greater than or equal to 2.306. In Figure 11.4, we see that the t-statistic is 3.77, which is greater than the critical t-value, so we can reject the null hypothesis.

Table 11.5 illustrates several important things about the t-statistic. First, note that all of the values in the table are positive, even though t-statistics can be negative. To use the table, one has to take the absolute value of the t-statistic. The only information that the sign adds is that it tells us which mean is greater. If the mean for Group 1 (male) is greater than the mean for Group 2 (female) the t-statistic will be positive. If the mean for Group 1 is less than the mean for Group 2 the t-statistic will be negative. Another thing to note from the table is

[5] Strictly speaking, target t-values associated with the critical region should be determined before the test statistic is calculated. Since computers do most this work for us all in one step, the order is not as important practically. However, the researcher should declare the acceptable alpha levels before conducting statistical analysis.

Table 11.5 Portion of the *t*-distribution. Degrees of freedom are associated with each row, while alpha levels (probability values) are associated with columns.

	Alpha Levels							
	One-tailed:	0.4	0.25	0.1	0.05	0.025	0.01	0.005
	Two-tailed:	0.8	0.5	0.2	0.1	0.05	0.02	0.01
Degrees of Freedom	1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
	2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
	3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
	4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
	5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
	6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
	7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
	8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
	9	0.261	0.703	1.383	1.833	2.262	2.821	3.250

that as sample size increases (as evident from the *df*), the critical *t*-value decreases, and so bigger sample sizes require smaller differences (also note the rate at which the critical *t*-values shrink as a result of increases in the *df*). Finally, keeping the *df* constant, notice how the critical *t*-value changes as a function of the alpha levels. To really understand how much easier it is to achieve significance with a one-tailed test instead of a two-tailed test (and why they carry a great risk of a Type I error), note that if we conducted a one-tailed test our critical *t*-value would have been 1.860 instead of 2.306.

The description above is for an independent samples *t*-test which arguably is the most common type of *t*-test conducted in IIR research. Another type of *t*-test which is also used in IIR research is the paired-samples *t*-test. To illustrate the difference between these two tests, we will re-visit the example study describing IL. In the original design of this study, the treatment group was compared to the population, but this study could have been designed in at least two other ways. In the first alternative design, the researcher could have two sample groups of subjects,[6] with one group receiving the education program and the other group not receiving the program. In the second alternative design, the researcher could have used a single group of subjects and given them a pre-test to elicit a baseline measure of IL, administered the program, and then given a post-test to measure IL. Neither of

[6] This design assumes that subjects are randomly assigned to conditions.

these designs requires knowledge of the population mean or standard deviation. Instead, to examine the null hypothesis in the first design alternative we would compare the mean IL scores of the two groups using the standard, independent samples t-test, while in the second design alternative we would compare the pre- and post-test IL scores of individual subjects. These two design alternatives illustrate the difference between when would use an *independent samples t-test* and when one would use a *paired-samples t-test*. The independent samples t-test examines differences in the means of two separate groups of subjects, while the paired-samples t-test examines differences within-subjects — subjects' pre- and post-test scores are compared with one another.

The formula for the paired-samples t-test is nearly identical to that used for the independent samples t-test except that the sample data are difference scores and are represented by D instead of X. This formula is

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}} \quad \text{where } s_{\overline{D}} = \sqrt{\frac{s^2}{n}}.$$

Notice that the *population mean difference* (instead of the population mean) is subtracted from *sample mean difference*. Also notice that the estimated standard error is given for the mean difference instead of the mean score. As was the case with the independent samples t-test, the *population mean difference* represents the null hypothesis and is set to 0 ($\mu_D = 0$).

11.2.2.3 *F*-statistic and Analysis of Variance (ANOVA)

Very often in IIR evaluations, researchers study a single independent variable with more than two groups or levels. Using the IIR research scenario, an example variable with three levels is system type. When one wants to compare the differences between three or more means, then *analysis of variance* (ANOVA) is used. There are several types of ANOVAs. In this paper, the single and multi-factor ANOVAs are discussed. The multivariate ANOVA (MANOVA) is introduced, but not discussed in detail.

The t-statistic measured the difference between sample means divided by the difference expected by chance (estimated standard error). The statistic produced by an ANOVA is called the F-ratio or F-statistic. It is similar to the t-statistic, expect it uses variance among groups, rather than means in its computation. Although variance is compared directly instead of means, the purpose of the test is to evaluate differences in means between conditions. Because there are more than two means, it is easier to compare variances.

The basic ANOVA formula is the ratio of differences in variances between the sample means to differences in variances expected by chance (called the *error variance*). For an ANOVA to be significant two major things need to happen: there needs to be a difference between at least one pair of means (for instance, between the mean performance for System A and System B, or System A and System C or System B and System C)[7] and the variance within each group must be small. When the within group variance increases, the error variance (the denominator in the F-ratio) also increases. The basic formula is given below.

$$F = \frac{variance\_between\_treatments}{variance\_within\_treatments}.$$

Although it is seemingly simple, it involves a number of calculations and uses the means and variances of the each group, as well as the total mean and variance.[8] When presenting the details of this formula, the example from the IIR research scenario stated above will be used, where system type is the independent variable and performance is the dependent variable. Thus, the researcher is interested in determining whether there is a statistically significant difference in performance according to system type.

To compute the ANOVA, the first step is to compute the overall variability using all of the data. After we have done this, we need to partition this variability into two components: variability between each condition and variability within each condition. This is known as *between-treatments variance* and *within-treatments variance*. The

[7] This is called *pair-wise comparisons*, which will be discussed in more detail later.

[8] In this discussion, we will use the term *grand* to refer to the total mean and variance of all of the subjects, and *sample* to refer to the means and variances for subjects in each condition (also called group or level).

between-treatments variance helps us understand how much variance occurs as a result of the different treatments, while the within-treatments variance helps us understand how much variance occurs by chance. Recall that the logic of hypothesis testing is to determine whether the probability of observing our results by chance is less than 5% (or 1%, depending on the alpha level). ANOVA attempts to measure this chance, or error variance.

Differences due to chance are typically attributed to two sources: individual differences and experimental error. These differences exist, but are independent of the treatment and should not be attributed it — the ANOVA attempts to partition these differences from differences caused by the treatment. Although steps can be taken to minimize the error introduced by these sources, such steps will never eliminate this error. For instance, random assignment to condition should distribute subjects with varying individual differences equally across groups, but there are lots of individual differences, so some error will still likely exist. Experimental error is related to how the experiment was conducted; a particular concern is the error introduced by poorly designed instruments, measures and protocols. A researcher can work to minimize these differences as well, but it is unlikely that they will be completely eradicated. To account for such chance differences, the ANOVA formula can be recast as

$$F = \frac{treatment\_effect + difference\_due\_to\_chance}{differences\_due\_to\_chance}.$$

The formula above helps us to not only understand the logic of the ANOVA, but also interpret the quotient it produces. In the above formula, if there was no treatment effect (treatment effect = 0) then the numerator and denominator would be equal since they measure the same thing, chance difference. This will result in an F-ratio of 1.00. When this happens, we fail to reject the null hypothesis since the only differences that were observed were due to chance. An F-ratio of 1 will never be statistically significant. It is also the case that F-ratios will always be positive numbers (unlike t-statistics). Since a value of 1 indicates no statistical significance, the values of the F-distribution will accumulate around this point on the distribution graph, similar to

how they accumulate around the mean in a normal distribution. Recall that to determine the significance of a z- or t-statistic we examine locations on a normal distribution. To determine whether an F-statistic is significant, the F-distribution is consulted. Instead of being normally distributed, the F-distribution is positively skewed. This is because F-values are always positive (essentially we are looking at the positive half of the normal distribution).

The exact shape of the F-distribution and the critical values needed to obtain statistical significance are determined by the alpha level and the degrees of freedom (*df*). There are three different *df* associated with an ANOVA: within-treatments (df_{within}), between-treatments (df_{between}) and total (df_{total}). The df_{total} is the sum of df_{within} and df_{between}, and is also equal to $n - 1$, where n is equal to the total sample size. The df_{within} is the difference between the number of levels of the independent variable and the total sample size. The df_{between} is equal to the number of levels of the independent variable (k) minus 1 ($k - 1$).

To demonstrate the relationship among sample size, levels of a variable, alpha levels, and critical F-values, a portion of the F-distribution is shown in Table 11.6 [27]. The df_{between} and df_{within} are used to locate F-values in this table. For instance, if we were examining the relationship between our three experimental IIR systems and performance, and we recruited nine subjects and randomly assigned three of these subjects to each system, then *df*(between) = 2, or 3 − 1 and *df*(within) = 6, or 9 − 3. Our critical F-values are equal to 5.14 for significance at the 0.05 alpha level or 10.92 for significance at the 0.01 alpha level.

In Table 11.6, notice the relationship between sample size (as evidenced by df_{within}) and critical F-values: as one increase the other decreases, but the rate at which this happens diminishes at some point. Also notice the extremely large F-values in the first row of the table. These represent cases where there is only one more subject than there are levels of the independent variable. In fact, to use this table, n must always be at least one point greater than the number of levels (k). For example, if a variable with eight levels was being tested on eight subjects (one subject per level) ($df_{\text{between}} = 7$ and $df_{\text{within}} = 0$), the critical F-value would be indeterminable. However, if nine subjects were used

Table 11.6 Portion of the F-distribution for $p < 0.05$ (top number in cell) and $p < 0.01$ (bottom number in cell). Degrees of freedom within groups is associated with rows, while degrees of freedom between groups is associated with columns.

	Degrees of Freedom$_{\text{between}}$						
Degrees of Freedom$_{\text{within}}$	1 ($p<0.05$) ($p<0.01$)	2 ($p<0.05$) ($p<0.01$)	3 ($p<0.05$) ($p<0.01$)	4 ($p<0.05$) ($p<0.01$)	5 ($p<0.05$) ($p<0.01$)	6 ($p<0.05$) ($p<0.01$)	7 ($p<0.05$) ($p<0.01$)
1	0161 4052	0200 4999	0216 5403	0225 5625	0230 5764	0234 5859	0237 5928
2	18.51 98.49	19.00 99.00	19.16 99.17	19.25 99.25	19.30 99.30	19.33 99.33	19.36 99.34
3	10.13 34.12	09.55 30.92	09.28 29.46	09.12 28.71	09.01 28.24	08.94 27.91	08.88 27.67
4	07.71 21.20	06.94 18.00	06.59 16.69	06.39 15.98	06.26 15.52	06.16 15.21	06.09 14.98
5	06.61 16.26	05.79 13.27	05.41 12.06	05.19 11.39	05.05 10.97	04.95 10.67	04.88 10.45
6	05.99 13.74	05.14 10.92	04.76 09.78	04.53 09.15	04.39 08.75	04.28 08.47	04.21 08.26

($df_{\text{between}} = 7$ and $df_{\text{within}} = 1$), the critical F-value is determinable, although impossibly large ($F = 237$). Thus, ANOVA is more accurate when there is a reasonable relationship between k and n. ANOVA is most accurate when there are equal sample sizes across condition. ANOVA is robust enough to handle unequal sample sizes, but the overall samples size should be relatively large and there should not be a huge discrepancy between the sample sizes for each condition.

ANOVA was originally developed for experimental situations where researchers have control over the assignment of subjects to conditions. However, in some cases ANOVA may be used to examine the effects of a variable that was not originally controlled in a study. For instance, in the IIR research scenario above, a researcher might be interested in examining the relationship between familiarity and performance. In this example, familiarity would be a *quasi-independent* variable since it was not manipulated by the researcher. It is unlikely that the distribution of familiarity scores will be equal across the five levels of the scale. If the distribution is too skewed, then the researcher might, for instance, consider reducing the five levels into three.

Conducting an ANOVA requires a number of calculations. First, recall that sample variance (which makes up both the numerator and denominator of the F-ratio), is equal to the sums of squared deviations (SS) divided by the *df*. We need three types of variances to compute the F-statistic: between- and within-treatment variance, as well as total variance. Thus, three different sums of squared deviations (SS) values must be computed, along with three *df* values. Once we have these values we can compute the F-statistic; the following formula which consolidates everything can be used

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}.$$

To demonstrate all of the computations, let us revisit our original example where a researcher is interested in examining the differences among three systems with respect to performance (Figure 11.1). The original description had each subject use each of the three systems (within-subjects design). However, the basic ANOVA assumes that observations are independent and in order to use the ANOVA formula used in this section, we would need to have an experimental design where each subject only used one of the systems (between-subjects design). Otherwise, we would compute a *repeated-measures ANOVA*. Figure 11.5 illustrates the computation of the F-statistic with sample data assuming that each subject only uses one system. This example is divided into four parts:

- Part (I) shows the computation of the *within groups SS*. Deviations are taken using the individual group means. The sums of these squared deviations are then summed to form the *within groups SS*.
- Part (II) demonstrates the computation of the *total SS*. Deviations are taken using the grand mean instead of the individual group means.
- Part (III) demonstrates the computation of the *between groups SS*. This is computed by taking the deviations between the grand mean and each individual group mean. These values are then squared and multiplied by the n of each group. These values are summed to form the *between groups SS*.

Research Question: What is the relationship between system type and performance?
Sample Hypothesis: Subjects will perform better with System C than with Systems A or B.

System and Mean Performance		Part I: Score Deviations from Group Means d_w	Part I: d_w^2	Part II: Score Deviations from Grand Mean d_g	Part II: d_g^2	Part III: Group mean deviations from the Grand Mean d_b	Part III: d_b^2	Part III: Group n times d_b^2
System A	0.2347	0.0525	0.0028	0.1121	0.0126			
	0.3426	-0.0554	0.0031	0.0042	0.0001			
	0.2788	0.0084	0.0001	0.0680	0.0046			
	0.3046	-0.0174	0.0003	0.0422	0.0018			
	0.2753	0.0119	0.0001	0.0715	0.0051			
Mean	0.2872					0.0596	0.0036	0.0178
n	5							
System B	0.3123	-0.0685	0.0047	0.0345	0.0012			
	0.1258	0.1180	0.0139	0.2210	0.0488			
	0.2338	0.0100	0.0001	0.1130	0.0128			
	0.3104	-0.0666	0.0044	0.0364	0.0013			
	0.2368	0.0070	0.0001	0.1100	0.0121			
Mean	0.24382					0.1030	0.0106	0.0530
n	5							
System C	0.5477	-0.0383	0.0015	-0.2009	0.0404			
	0.4878	0.0216	0.0005	-0.1410	0.0199			
	0.5238	-0.0144	0.0002	-0.1710	0.0313			
	0.5011	0.0083	0.0001	-0.1543	0.0238			
	0.4866	0.0228	0.0005	-0.1398	0.0195			
Mean	0.5094					-0.1626	0.0264	0.1232
n	5							
Grand Mean	0.3468	**Within *SS*** (Σd_w^2)	0.0324	**Total *SS*** (Σd_g^2)	0.2354		**Between *SS*** ($\Sigma n d_b^2$)	0.1940

Part IV

Summary table

Source	*SS*	*d.f.*	*MS*	*F*
Between	0.1940	2	0.0970	35.9259
Within	0.0324	12	0.0027	
Total	0.2264	14		

$$F = \frac{0.0970}{0.0027}$$

= 35.9259

Fig. 11.5 Computation of the F-statistic using sample performance data for three systems (A, B and C).

- In Part (IV), the *MS* values are computed as the *SS* divided by the *df*, so that $MS_{\text{between}} = SS_{\text{between}}/df_{\text{between}}$. The quotients are then used to compute the F-statistic.

Recall that for our example, the critical F-values are equal to 5.14 for significance at the 0.05 alpha level or 10.92 for significance at the 0.01 alpha level. Thus, our F-statistic of 35.9259 is significant at the $p < 0.01$ level. Although our ANOVA is statistically significant, we do not know between which pairs of systems there were significant differences.

For instance, mean performance with System A might be significantly different from mean performance with System B and System C, but there might not be any significant difference between System B and System C. There are actually three pair-wise comparisons that we need to make — System A, System B; System A, System C; System B, and System C. To evaluate these pair-wise differences, *post-hoc* tests are conducted. A number of *post-hoc* tests can be used, including *Scheffé*, *Tukey HSD*, and *Bonferroni*. The difference among these tests is beyond the scope of this paper, but the *Scheffé* test is one of the safest and most conservative tests. Using a safe test (and by safe test it is meant a test that reduces the risk of a Type I error) is particularly important with *post-hoc* analysis because of the number of pair-wise comparisons being made. Essentially, one is conducting hypothesis testing for each pair. As one does more tests, the risk of a Type I error accumulates. This is called *experiment-wise alpha level*. The basic notion is that conducting more tests increases the risk that a statistically significant result will happen just by chance. In fact, this is one reason why it is better to conduct an ANOVA instead of multiple *t*-tests: the more tests you conduct, the greater the chance of a Type I error. What makes the *Scheffé* test conservative is that during the pair-wise comparisons, the between-treatments *df* from the ANOVA is used, even though only two groups are being compared. Thus, there are fewer *df*, which makes the test harder. The implication of this is that it is possible to have a statistically significant ANOVA, but no statistically significant *post-hoc* tests.

11.2.2.4 More ANOVAs

In the example above, we explored the effect of one independent variable (system type) on one dependent variable (performance). The basic one-way ANOVA described above also accommodates situations where a researcher examines a single independent variable in relation to multiple dependent variables. Sometimes researchers are interested in looking at more than one independent variable in relation to a single dependent variable. In this case, a *multi-factor ANOVA*[9] is conducted. When the researcher is interested in examining more than

[9] This is also called a *uni-variate ANOVA*, where *uni* refers to the dependent variable.

		System			
		A	B	C	Sex $\overline{X}$
Sex	Males	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$
	Females	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$
	System $\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$	Grand $\overline{X}$

Fig. 11.6 Example of a 3×2 factorial design using system type and sex. Numbers in cells represent means (of, for example, performance).

one independent variable in relation to more than one dependent variable, a *MANOVA*[10] is used.

Earlier in this paper, the factorial design was introduced. Studies that are designed in this way are typically appropriate for multi-factor ANOVAs. In fact, such representations are useful for understanding what is being compared in the multi-factor ANOVA and what types of computations are required. Figure 11.6 presents a 3×2 factorial representation of a relationship that could be studied from the IIR research scenario. In this example, the impact of two variables, *system type* and *sex*, on *performance* is investigated. Note that system type is an independent variable, since it was manipulated by the researcher, sex is a quasi-independent variable and performance is a dependent variable. (From this point forward, these independent variables will be called *factors* to coincide with the language of factorials.)

The impact of each one of these factors on the dependent variable is called a *main effect*. For instance, system type may have an impact on performance (e.g., all subjects perform better with System A, than Systems B or C, regardless of sex), and sex may have an impact on performance (e.g., female subjects perform better than male subjects regardless of system). In this example, there are two possible main effects. The number of possible main effects is equal to the number of factors.

It may also be the case that system type and sex interact. For instance, females may perform better with System A than System B or C, but males may perform better with System B than System A or C. This is called an *interaction effect*. The purpose of a multi-factor

[10] This is also called a *multi-variate ANOVA* where *multi* refers to the dependent variable.

ANOVA is to test the main effects and the interaction effects.[11] The computation behind a multi-factor ANOVA is nearly identical to that of a single factor ANOVA, expect that F-statistics are computed for each possible relationship (one for each factor and one for all possible interactions). Most multi-factor ANOVAs involve two or three factors. Anything beyond that generates a large number of possible main effects and interaction effects and such results can be extremely difficult to understand and interpret. Moreover, designs with larger numbers of factors require larger numbers of subjects.

Figure 11.6 is useful for conceptualizing the comparisons made during a multi-factor ANOVA. First, recall that ANOVA concentrates on variances as a way to determine whether differences between means are significant. The means are not compared directly, although they are used to compute *SS*. In Figure 11.6, the ANOVA testing for a main effect for sex would compare the row totals, the ANOVA testing for a main effect for system would compare the column totals and the ANOVA testing for the interaction would compare the values in the cells of the table. Thus, we can talk about row means, columns means, cell means, and the overall (grand) mean. Each of these values also has a variance associated with it, which is what is compared with the ANOVA. To calculate the F-ratio for the interaction, the ANOVA first identifies differences that cannot be explained by the main effects. These extra differences are then evaluated to determine whether there is a significant interaction effect. The entire computation of the multi-factor ANOVA will not presented here, because it involves a large number of steps. Instead, means and standard deviations for a sample of 30 subjects are added to the above example and presented in Figure 11.7.

In this example, there are significant main effects for both system and sex: females performed significantly better than males [$F(1,30) = 12.46,\ p < 0.01$)] and subjects performed significantly better with System C than with System A or B [$F(2,30) = 5.55, p < 0.01$)].[12] There

[11] A researcher is not required to have hypotheses about all possible effects. One may only be interested in the interaction effect, but not the main effects.

[12] We would technically need a *post-hoc* test to determine that the difference was C > A, B, but a visual inspection of the means suggests that this would be the only significant relationship.

		System			
		A	B	C	Sex $\overline{X}$
Sex	Males	0.336 (0.023)	0.132 (0.123)	0.214 (0.214)	0.227 (0.088)
	Females	0.130 (0.012)	0.338 (0.024)	0.334 (0.334)	0.267 (0.107)
	System $\overline{X}$	0.233 (0.110)	0.235 (0.110)	0.274 (0.077)	0.247 (0.099)

Fig. 11.7 3×2 factorial with one independent variable, system type, and one quasi-independent variable, sex. The values in the cells represent sample means (standard deviations) for performance ($n = 30$).

is also a significant interaction effect $[F(1,30) = 122.59,\ p < 0.01)]$: males performed best with System A, second best with System C and worst with System B (A > C > B), while females performed best (and about the same) with System B and C and worst with System A (B, C > A). We would need to conduct *post-hoc* tests to pinpoint between which pairs the significant differences occurred.

There are several other types of ANOVAs that will not be discussed in detail. The underlying formulas for computing these ANOVAs are similar to those presented above, but require more computations because there are more comparisons. *MANOVA* (Multiple Analysis of Variance) is basically a combination of the single factor and multi-factor ANOVAs discussed above in that it is used when the researcher is examining the effects of multiple factors on multiple dependent measures. There is also a special version of ANOVA that deals with between-subject independent variables. This is called repeated-measures*ANOVA*. Finally, g*eneralized linear modeling* (GLM) allows one to develop a function describing the relationship among the independent and dependent variables based on significant ANOVA results. This is similar in nature to linear regression, which is discussed briefly below.

11.2.2.5 Measures of Association

Another set of statistical techniques that are often used in IIR evaluations is correlation. There are many kinds of correlation coefficients including Pearson's r, Spearman's *rho*, and Kendall's *tau*. Correlation coefficients are measures of association; basically these coefficients

describe how scores on two variables co-vary.[13] For example, if a subject has a high performance score is the subject likely to give the system a positive usability rating? One important distinction to keep in mind is that correlation does not show causality. Correlation only shows that two things co-vary. Performance and age might be correlated, but this does not mean that performance causes a person's age or that age causes a person's performance. It only means that these things are systematically related. In this paper, we will look at Pearson's r and Spearman's *rho*. Kendall's *tau* is also used a lot in IIR and IR more broadly. Spearman's *rho* and Kendall's *tau* are used to test similar kinds of relationships. In the interest of space, only Spearman's *rho* is presented. Spearman's *rho* is technically a non-parametric statistic, so it will be presented in another section.

Correlation coefficients vary between -1 and $+1$. The magnitude of the coefficient indicates its strength, while the sign indicates if the relationship is positive or negative. A positive relationship indicates that increases in one variable are associated with increases in the other variable (or, conversely, that decreases in one variable are associated with decreases in the other variable). A negative relationship indicates that increases in one variable are associated with decreases in the other variable. This is also known as an inverse correlation. A value of zero indicates that there is no relationship between the two variables, while $+1$ and -1 indicate functional relationships.

Magnitude is very important for interpreting the meaningfulness of the correlation coefficient. It is possible (and common) to find statistically significant correlation coefficients that are actually quite weak, so one should always pay attention to the value of the coefficient. The real problem is that the correlation coefficient does not actually represent the accuracy with which predictions can be made. For instance, a correlation coefficient of 0.30 does not mean that given one variable the other variable can be predicted 30% of the time or with 30% accuracy. The strength of the relationship lies in the squared correlation coefficient (r^2), so that a correlation of 0.30 means that given one variable

[13] Correlation coefficients can be computed for more than two variables, but in this paper we will just consider the relationship between two variables.

Table 11.7 Cohen's and Guilford's guidelines for interpreting correlations.

Cohen (1988)	Guilford (1956)
0.10–0.29 small	< 0.20 slight, almost negligible
0.30–0.49 medium	0.21–0.40 low
0.50 + large	0.41–0.70 moderate
	0.71–0.90 high
	> 0.91 very high

the other variable can be predicted with 9% (0.30^2) accuracy. Clearly, this gives a very different view of the strength of coefficient. For correlation coefficients below 0.50, differences between the actual coefficient and the r^2 values are quite pronounced. Thus, one should be extremely cautious interpreting any statistically significant correlation coefficients, especially those whose values are small.

There are several guidelines for interpreting the magnitude of a correlation. Two such interpretations are given in Table 11.7. Both authors stress that these are guidelines, rather than absolutes. Guildford [113] offers more distinctions than Cohen [56]. Both use the absolute values of the coefficient.

Pearson's r

The Pearson's correlation (r) is one of the most common correlation coefficients. Traditionally, it is used with continuous data types (interval or ratio level data) and measures linear relationships. The calculation for Pearson's r examines the degree to which two variables vary together in relation to the degree to which they vary separately. The formula for Pearson's r is given in Figure 11.8, along with an example from the IIR Research Scenario which looks at the relationship between query length and performance. To calculate Pearson's r, we use the *sum of products of deviations*, which is similar in nature to the sum of squared deviations calculation that was used in the t-test and ANOVA.

The sum of products of deviations is illustrated in Figure 11.8. For any given subject with scores on variables X and Y, the deviations of each of these scores from their respective sample variable means are multiplied. After this is done for each subject, these values are summed to form the sum of products of deviations. This value represents the

	Raw Data		Deviations from Means				
Subject	X (Query Length)	Y (Performance)	x	y	xy	x^2	y^2
1	2	0.2445	−0.5000	−0.0219	0.0110	0.2500	0.0005
2	3	0.3022	0.5000	0.0458	0.0229	0.2500	0.0021
3	4	0.3387	1.5000	0.0723	0.1085	2.2500	0.0052
4	1	0.1804	−1.5000	−0.0860	0.1290	2.2500	0.0074
5	2	0.2556	−0.5000	−0.0108	0.0054	0.2500	0.0001
6	3	0.3433	0.5000	0.0769	0.0385	0.2500	0.0059
7	3	0.2990	0.5000	0.0326	0.0163	0.2500	0.0011
8	4	0.3711	1.5000	0.1047	0.1571	2.2500	0.0110
9	2	0.1915	−0.5000	−0.0749	0.0375	0.2500	0.0056
10	1	0.1277	−1.5000	−0.1387	0.2081	2.2500	0.0192
Σ	25	2.664			$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$
Mean	2.50	0.2664			0.7343	10.5000	0.0581

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2.}} = \frac{0.7343}{\sqrt{10.50 \cdot 0.0581}} = \frac{0.7343}{\sqrt{0.6101}} = \frac{0.7343}{0.7811} = +0.9401$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9401\sqrt{8}}{\sqrt{1-0.8838}} = \frac{2.6590}{\sqrt{0.1162}} = 7.8000$$

Fig. 11.8 Computation of Pearson's r with sample query length and performance data.

extent to which the two variables co-vary. To calculate the extent to which the two variables vary separately, each individual X and Y deviation score is squared and then summed within each variable. These values are then multiplied and square-rooted. This calculation is similar to the variance measures used in the previous statistics.

The Pearson's r for our example data yields a value of +0.9401, which we can see is quite large in magnitude. This suggests that there is a strong positive correlation between query length and performance and the r^2 is 0.8838 which demonstrates that query length explains quite a bit of the variability in performance.

Although our coefficient looks strong, we still need to evaluate it with respect to probability. The null hypothesis states that there is no correlation between the two variables, in which case we would expect $r = 0$. (Even when no relationship exists, the correlation coefficient is usually not zero. In most cases, it would be some non-zero value.) To evaluate the statistic, we need to compute a corresponding t-statistic and use the t-distribution to evaluate the likelihood of observing the coefficient. This formula is given in the bottom of Figure 11.8. This

requires knowledge of the Pearson's r, r^2 and the sample size. In the numerator, the r is multiplied by the square-root of the *df* $(n-2)$ for the sample. The *df* is also used to enter the t-distribution (along with the alpha level). Using Table 11.5, we can see that our critical value is 2.306. Our t-statistic is 7.80, which is significant at the 0.01 level, so we can reject the null hypothesis. Although the t-statistic is necessary in determining whether the null hypothesis can be rejected, many people do not realize that this is computed as part of correlation testing. It is not necessary to report this statistic; instead, we would report $r = 0.9401$, $p < 0.01$.

11.2.2.6 Regression

Correlation coefficients measure the extent to which two variables covary, but they do provide information about how to predict one value from another. Regression can be used to discover the function describing the relationship among two or more variables. Regression is a sophisticated set of statistical procedures and a discussion of these procedures is beyond the scope of this paper. There are several forms of regression including techniques for both linear and non-linear relationships, and for different data types. Regression is also useful for evaluating the importance of a set of predictor variables and determining which are the most useful for predicting a particular output variable.

11.2.2.7 Effect Size

Effect size measures the strength of a test statistic. Very often a researcher might want to go beyond just saying that statistically significant relationships were found, to a discussion and comparison of the strengths of such relationships. As was shown with the correlation coefficient, even when statistically significant relationships are found, they are not always meaningful. The value given by r^2 is a measure of *effect size*. It indicates the proportion of variability in one variable that can be determined by its relationship with another variable. Stated another way, it shows how much variance in one variable is explained by differences in another variable. In addition to r^2, there are several

other measures of effect size that can be used in conjunction with *t*-tests and ANOVAs. Two such measures will be presented here, *Cohen's d* for *t*-statistics and eta^2 for *F*-statistics. The interpretation of the values are taken from Cohen [56], who is very cautious about associating values with specific qualitative labels such as *small* and *large*. The use of Cohen's [56] interpretations of effect sizes are standard across a range of behavioral science disciplines and despite Cohen's cautiousness they provide useful heuristics for interpreting effect sizes.

Cohen's *d* can be computed in a number of ways, but the easiest is given below in the first formula, which uses the value of the *t*-statistic and the *df*. This formula assumes that sample sizes of the two groups that are being tested are equal. In cases where this is violated, another version of the formula can be used, which accounts for the differences. This version of the formula (b) includes values for the size of each sample (e.g., if males and females were being compared, one of these values would correspond to the number of male subjects while the other would correspond to the number of female subjects). Typical interpretations of Cohen's *d* are: small = 0.2, medium = 0.5 and large = 0.8.

$$d = \frac{2t}{\sqrt{df}} \quad \text{(a)} \qquad d = \frac{t\,(n_1 + n_2)}{\left(\sqrt{df}\right)\left(\sqrt{n_1 n_2}\right)} \quad \text{(b)}$$

(a) for equal sample sizes or (b) for unequal sample sizes

The computation for eta^2 is also relatively straight-forward: it is the ratio of the between-treatments SS to the total SS. These values are interpreted on a slightly different scale from Cohen's *d*: small = 0.10, medium = 0.25, and large = 0.40.

$$\eta^2 = \frac{SS_{bg}}{SS_T}.$$

11.2.2.8 Non-Parametric Tests

Unlike parametric tests, non-parametric tests make few assumptions about the distribution of variables in the population. Specifically, these tests do not rely on the assumption that variables are distributed normally in the population. Non-parametric tests are also useful for

analyzing discrete data types, such as nominal and ordinal measures. Non-parametric tests can also be considered as more robust since they make fewer assumptions than parametric tests and can be used in more situations. However, it is important to note that in general, non-parametric tests are not as sensitive as parametric tests and thus, the risk of Type II errors are greater. The important thing is for researchers to select the most appropriate test to ensure the credibility and integrity of their results rather than the significance.

There are a number of non-parametric tests that have been used in IIR including the *Mann-Whitney* test, *Wilcoxon Signed-Rank* test, *Kruskal-Wallis* test, *Spearman's Rho*, and *Chi-square*. We will look closely at *Spearman's Rho* and *Chi-square* because they test different data types and relationships than any of the previously discussed tests. The other non-parametric tests are not discussed since they offer non-parametric alternatives to other tests. The *Mann-Whitney* and *Wilcoxon Signed-Rank* tests offer alternatives to the *t*-test, while the *Kruskal-Wallis* test offers an alternative to ANOVA.

Spearman's rho

Spearman's *rho* is correlation coefficient which has been typically used to evaluate ordinal data and to test for relationships that are not necessarily linear. Thus, the Spearman correlation measures the consistency of the relationship between two variables, but it does not say anything about its form. Ordinal level data is often rank-level data. For instance, subjects in a study might be rank-ordered from the best performer to the worst performer. It was noted much earlier in this paper that the Likert-type scale data that is common in IIR evaluations is technically ordinal level data, although it is promoted to interval level status so that more sophisticated analysis can be performed with it. However, when exploring correlation, it is possible to study this data at its native level using Spearman's *rho*.

As a reminder, ordinal level data tells us that one thing is [better or worse] or [more or less] than another, but it does not tell us how much since the distances between points are not constant. This can be easily illustrated with the example above, where subjects are to be ranked according to how well they perform. Ratio level data (performance

score) could be used to create this ranking, but once the ranking was done, we would only know that Subject A was 1, Subject F was 2, Subject B was 3, etc. We would not know how much better Subject A was than Subject F, and there would be no guarantee that the difference in performance scores between Subject A and F was equal to the difference between Subject F and B. Thus, some information is lost when converting ratio level data into ordinal level data.

The calculation of Spearman's *rho* is displayed in Figure 11.9, along with sample data from the IIR Research Scenario that investigates the relationship between familiarity and usability. This formula is actually a simplification of the Pearson's r formula that assumes that scores are ranked. Thus, before using this formula, the raw data must be converted into ranked data. If the original measure is ranked data (e.g., subject ranking according to performance) then this step is not necessary. The data in the example come from two scales and so the scores first need to be transformed into rank values. Although this is not shown in the

	Raw data		Ranking values		Differences	
Subject	X (Familiarity)	Y (Usability)	x_r	y_r	D $(x_r - y_r)$	D^2
1	4	5	7.5	9.5	–2	4
2	1	1	1.5	1.5	0	0
3	3	4	5	6.5	–1.5	2.25
4	1	4	1.5	6.5	–5	25
5	3	1	5	1.5	3.5	12.25
6	3	2	5	3.5	1.5	2.25
7	2	4	3	6.5	–3.5	12.25
8	4	2	7.5	3.5	4	16
9	5	5	9.5	9.5	0	0
10	5	4	9.5	6.5	3	9
						$\Sigma = 83$

Formula:

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$
$$= 1 - \frac{6(83)}{10(100 - 1)}$$
$$= 1 - 0.503$$
$$\rho = 0.497$$

Fig. 11.9 Calculation of Spearman's *rho* using sample familiarity and usability data.

example, the easiest way to do this is to order the scores from smallest to largest and assign rank values to each position. In our example data, there are a few ties — e.g., four subjects used a usability rating of four. When two or more scores are tied, the mean of their ranked positions is computed and assigned to all scores with this value. In the example, two subjects used a usability rating of one. The corresponding rank value for these two subjects is 1.5, or (1+2/2). The differences between each subject's ranked X and Y scores are then computed and squared. The set of differences are then summed. This value is multiplied by six and forms the nominator of the *rho* formula. The denominator is a simple calculation using the sample size.

To determine whether the coefficient is significant and the null hypothesis can be rejected, the t-statistic formula displayed in Figure 11.8, along with the t-distribution is used. It is important to note that the Spearman's *rho* formula loses some of its accuracy when there are a lot of ties; if this is the case, then the researcher might want to explore an alternative coefficient.

Chi-Square

The chi-square test is used to compare the distribution of scores across two or more levels. Figure 11.10 illustrates some sample data that corresponds to the IIR Research Scenario where the researcher asks subjects to indicate which system they liked best. The numbers in this Figure represent the frequency of subjects who selected a particular system as their favorite. Is there a significant difference with respect to which system subjects prefer?

Since we are only looking at the distribution of one variable, our test is for *goodness of fit*, which examines how good the data fit the

	System			
	A	B	C	Total
Observed frequencies (O)	1	1	13	15
Expected frequencies (T)	5	5	5	15

$$\chi^2 = \sum \left[\frac{(O-T)^2}{T} \right] = \frac{(1-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(13-5)^2}{5} = 19.2$$

Fig. 11.10 Computation of χ^2 for sample system preference data.

distribution specified by the null hypothesis. The null hypothesis states that there are no differences in subjects' system preference. This is represented by the second row of Figure 11.10 — these values are referred to as the expected frequencies. If the null hypothesis is true, then the preference distributions should be roughly equal across system. Unless otherwise specified, the null hypothesis assumes the distributions will be equal across category. However, if it is known that the distributions in the population are unequal, then the expected frequencies can be adjusted. As mentioned earlier in the discussion of z-statistics, it is rarely the case in IIR that we know anything about the population, so the default null is almost always used. The purpose of the chi-square test is to compare the observed distributions with the null expectation. The alternative hypothesis, in this situation, simply states that the population is not divided equally among the various categories.

The formula for, and computation of, chi-square is shown at the bottom of Figure 11.10. This formula is equal to the sum of the squared differences between the observed and expected frequencies divided by the expected frequency. This formula basically measures the discrepancy between the observed frequencies and the expected (or theoretical) frequencies. The value of the chi-square statistic is directly related to the size of the discrepancy — the larger the discrepancy, the larger the chi-square value.

Similar to ANOVA, the chi-square distribution is positively skewed — the majority of scores will cluster around 0–1 — these values represent the null hypothesis. A statistically significant chi-square value will be out in the tail of the distribution. As with all statistical tests, the chi-square statistic also has an associated *df*, which is equal to the number of categories minus 1 $(C - 1)$ (in the example $3 - 1$). This value, along with an alpha level, allows one to enter the table of values that correspond to the chi-square distribution (Table 11.8) [27] to determine if a particular chi-square value is statistically significant. Table 11.8 tell us that the chi-square value for our sample data is beyond the critical value of 5.99 and is therefore, statistically significant. In fact, our chi-square statistic is significant at $p < 0.001$. We would report this as $\chi^2(2) = 19.2$, $p < 0.001$.

Table 11.8 Chi-square distribution.

	Probability values						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.3233	2.7055	3.8414	5.0238	6.6349	7.8794	10.828
2	2.7725	4.6051	5.9914	7.3777	9.2103	10.5966	13.816
3	4.1083	6.2513	7.8147	9.3484	11.3449	12.8381	16.266
4	5.3852	7.7794	9.4877	11.1433	13.2767	14.8602	18.467
5	6.6256	9.2363	11.0705	12.8325	15.0863	16.7496	20.515
6	7.8408	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550	26.125
9	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893	27.877

The chi-square test can also be used to test for independence when the distributions of two variables are being compared. Using our IIR example, we might examine whether there is a relationship among system preference and sex. This is similar in nature to correlation in that each subject has a value on two variables (system preference and sex) except that chi-square examines the frequency distributions since these variables are nominal. The null hypothesis in this case states that the distribution of system preferences will be the same for females as for males, or put another way, the frequency distributions will have the same shape for both females and males.

Some sample data is presented in Figure 11.11, along with the chi-square formula and the chi-square computation for our sample data. Note that when two variables are involved, the expected frequencies are

	System			Row Totals
	A	B	C	
Males:				
Observed Frequencies (O)	7	26	7	40
Expected Frequencies (T)	8	19	13	40
Females:				
Observed Frequencies (O)	7	7	16	30
Expected Frequencies (T)	6	14	10	30
Column totals	14	33	23	70 (Grand total)

$$\chi^2 = \sum \left[\frac{(O-T)^2}{T} \right] = \frac{(7-8)^2}{8} + \frac{(26-19)^2}{19} + \frac{(7-13)^2}{13} + \frac{(7-6)^2}{6} + \frac{(7-14)^2}{14} + \frac{(16-10)^2}{10} = 12.74$$

Fig. 11.11 Calculation of χ^2 using sample system preference and sex data.

a function of the characteristics of the sample along these two variables. For instance, in the sample there are 40 males and 30 females. Thus, the distribution of system preferences cannot be equal in terms of an absolute value, but must be equal proportionately. To compute the expected frequencies, first the proportions of subjects selecting each system are computed: System A ($14/70 = 20\%$), System B ($33/70 = 47\%$) and System C ($23/70 = 33\%$). The null hypothesis assumes that these same proportions will be observed for both males and females: Males [System A ($0.20 * 40 = 8$), System B ($0.47 * 40 = 19$), System C ($0.33 * 40 = 13$)] and Females [System A ($0.20 * 30 = 6$), System B ($0.47 * 30 = 14$), System C ($0.33 * 30 = 10$)]. While the computation and interpretation of the chi-square test for independence is the same as that for goodness of fit, the calculation of *df* differ. For tests of independence, $df = (R - 1)(C - 1)$, where R = number of rows and C = number of columns. Using this value in conjunction with the chi-square distribution in Table 11.8, we see that the critical value is 5.99 with alpha = 0.05. Our chi-square statistic is well beyond this and is even significant at the 0.005 level, so we can reject the null hypothesis. We would report this as $\chi^2(2) = 12.74$, $p < 0.005$.

11.2.3 Cohen's Kappa

The final statistic that will be presented is Cohen's Kappa, a measure of inter-rater reliability. Inter-rater reliability (also known as inter-coder reliability) shows the extent to which two or more people agree on how to classify a set of objects. It can be used to check the reliability between relevance assessments made by two people and to check the reliability of how a researcher has analyzed and classified qualitative data. Inter-rater reliability measures provide a much stronger measure of rating consistency than simple percent agreement since they take into account the distribution of responses and the amount of agreement that would happen by chance. Percent agreement is an inflated index of agreement and can be especially misleading when the underlying distributions are skewed.

Cohen's Kappa is not an inferential statistic. Instead, it produces a value similar to a correlation coefficient. The values for Kappa

range from 0–1.00, with larger values indicating better reliability. Most researchers accept Kappa values greater than 0.70 as satisfactory. If the value is less than this, then researchers will often revise classification rules, solicit more raters to apply these rules and then re-assess the statistic. Thus, an important part of this exercise is ensuring that the rules for classification are clear, can be easily used to distinguish between objects, and can be understood and executed by multiple people. Raters do not have to necessarily agree with the classification rules they just have to execute them consistently.

The formula for Cohen's Kappa is given below. To execute this formula, one should first build a contingency table displaying the ratings made by the raters in relation to one another. The diagonal of this table will show the total agreements made by the two raters, while the off-sets will show the disagreements. Row and column totals should be computed as well as expected frequencies (*ef*) for each classification category:

$$\mathrm{K} = \frac{\Sigma a - \Sigma ef}{n - \Sigma ef}$$

where Σa = sum of the agreements (diagonal), n is the total number of objects and

$$ef = \frac{row\_total \cdot col\_total}{overall\_total}$$

The computation of Cohen's Kappa is shown in Figure 11.12. Imagine that two raters have used a four point scale to classify the relevance of 259 documents. In total, the raters have agreed on 167 ratings (see diagonal). A simple percent agreement would show that the raters have agreed 64% of the time. However, the Cohen's Kappa shows the inter-rater agreement to be 12% points less (52%). Since this value is less than the target agreement of 70%, it would be necessary to refine classification rules and perform the ratings again. An examination of the disagreements can help identify where raters are having the most problems. In the example, most disagreements happen between relevance categories two and three. Thus, refining rules for distinguishing among these two types of relevance is the best place to start with the revisions.

Contingency Table

		Rater 1				Row
		1	2	3	4	Totals
Rater 2	1	70	7	5	0	82
	2	5	25	28	2	60
	3	6	22	32	5	65
	4	2	4	6	40	52
Column Totals		83	58	71	47	259

Expected Frequencies (*ef*)

$$ef_1 = \frac{82 * 83}{259} = 26.28$$

$$ef_2 = \frac{60 * 58}{259} = 13.44$$

$$ef_3 = \frac{65 * 71}{259} = 17.82$$

$$ef_4 = \frac{52 * 47}{259} = 9.44$$

$$\Sigma ef = 66.98$$

$$\Sigma a = 167$$

Cohen's Kappa

$$K = \frac{167 - 66.98}{259 - 66.98}$$

$$= \frac{100.02}{192.02}$$

= .52 (52%)

Fig. 11.12 Calculation of Cohen's Kappa using sample data from two raters, who have labeled a set of documents according to four levels of relevance.

11.2.4 Statistics as Principled Argument

The title of this section takes its name from the book by Abelson [1], who describes five properties that determine the persuasiveness of a statistical argument: magnitude, articulation, generality, interestingness and credibility. These properties emphasize that statistics assist with analysis, but their messages have to be interpreted by humans. Abelson is careful to point out that single studies are not definitive, significance tests provide limited information, and interpretation of statistical results is just as important as the statistics themselves.

Magnitude is related to the strength of a result. Even in cases where statistical significance is found one must look critically at the strength of the relationship. In previous sections, it was shown that some statistically significant correlation coefficients are not meaningful. Calculations of effect size are the most common methods for assessing magnitude. If effect sizes are small, then researchers should be more conservative with their interpretations and conclusions.

Articulation is the degree of specificity with which results are presented. An example of poor articulation is when a researcher conducts

an ANOVA to test for differences between groups, but does not conduct follow-up tests to pinpoint differences. The main point about articulation is to understand that what is being studied is typically complex and it is sometimes necessary to look closely at individual cases to understand what is happening rather than relying on the overall statistic.

Generality is the applicability of study results and conclusions to other situations. Researchers typically use reductive methods to examine very narrow problems, which can impact the generality of the results. Abelson advocates for a wide range of investigations that center on the same phenomena. The implications are that a single study should be one of a larger body of research designed to investigate a particular problem. Single study results provide support for particular conclusions, but not definitive conclusions. The accumulation and analysis of data from many different studies designed to investigate the same problem enhance generality.

Interestingness and credibility are attributes of the research story in which statistical arguments are placed. Abelson [1, p. 13] adopts the viewpoint that "for a statistical story to be theoretically interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue". Abselson [1, p. 13] notes that the importance of an issue contributes to its interestingness, where importance is defined as "the number of theoretical and applied propositions needing modification in light of the new result". The criterion of importance is typically used to evaluate the quality of research and good writing practice dictates that authors include some statement of importance in their research reports.

The credibility of research, according to Abelson, should be evaluated according to the soundness of the method and theoretical coherence. The soundness of the method is a quality of research design and data analyses. One of the most common problems in IIR reports is that researchers do not provide enough detail for readers to evaluate the credibility of the method. Without being able to understand the experimental design and procedures, it is nearly impossible to assess the credibility of the results. Credible results depend on a credible research design.

Theoretical coherence is a bit more difficult to discuss in the context of IIR research since at present much of the research is not theoretically based and there is a strong underlying current of applied science rather than basic science. However, if a research claim contrasts with prevailing theory or belief, then the researcher should be prepared to rule out alternative explanations of the findings and demonstrate why the alternative explanation is the most parsimonious. The researcher must also demonstrate the coherence of an alternative theory by showing that it can explain a number of interconnected findings. The burden of proof ultimately lies with the researcher and having statistically significant results does not provide proof, only evidence (provided the method is sound).

12

Validity and Reliability

Validity and reliability assessments can be applied to both the method used to conduct the study as well as to specific measures. Validity is the extent to which methods and measures allow a researcher to get at the essence of whatever it is that is being studied, while reliability is the extent to which the method and measures yield consistent findings. Validity and reliability assessments are particularly important for understanding the overall quality and limitations of a study, and ultimately, the extent to which research results are believable and generalize. All studies can be critiqued in terms of validity and reliability and no study will be free of validity and reliability issues. There is a tension between validity and reliability, so optimizing both of these in a single study is usually not possible. Although measures are technically part of the method, these two concepts will be discussed separately, since there are special validity and reliability issues related to measurement. Method will be used to refer to the specific procedures used to conduct the study and measures will refer to instruments and metrics.

There are two broad validity classes: internal validity and external validity. Internal validity is related to the quality of what happens during the study. One of the most common threats to internal validity is

instrumentation which is related to the quality of the instruments and measures. If an instrument yields poor or inaccurate data, then the results of the study are unlikely to be valid. External validity is the extent to which results from a study generalize to the real world. A study may have good internal validity, but the results may be meaningless outside the particular experimental situation. Thus, internal validity is a necessary but not sufficient condition for external validity.

Certain methods are associated with certain levels of validity and reliability, regardless of how they are executed. In addition, each individual study will have validity and reliability issues that are specific to how that particular study is executed. Laboratory studies are generally thought to be less valid, but more reliable than naturalistic studies. Laboratory studies typically involve artificial situations that are tightly controlled by the researcher. As a result, it is questionable whether the behavior exhibited by subjects in a laboratory study is the same as the behavior they would exhibit in a natural environment. However, with a detailed study protocol and certain number of controls, it is possible for each subject to experience the evaluation situation in a similar way, so laboratory studies are generally characterized as having high reliability. Naturalistic studies provide a more realistic view of subjects' searching behaviors, but such studies are not typically controlled and so it is impossible to ensure that each person experiences the study in the same way.

Demand effects and reactivity are important concerns in studies with human subjects. The research context can demand that subjects behave in particular ways. This includes who administers the experiment, where it is administered and how it is administered. One specific type of demand effect is experimenter demand effects where the experimenter (either consciously or unconsciously) communicates to subjects how they should behave. The experimenter, in most cases, knows a lot more about the study than the subjects, including the desired outcome. Thus, there is a danger that this knowledge is communicated either implicitly or explicitly during the experimenter–subject interactions. In some research, safeguards are put in place to protect against this. For instance, in medical research double-blind experiments are common. In this situation, neither the subject nor researcher knows

who is assigned to the experimental condition. In other disciplines, it is customary for those administering the study to be ignorant of the goals, objectives and hypotheses of the study. It is also the case that subjects might try to guess the purpose of the study and act accordingly to please the experimenter even if this does not reflect their behaviors and desires.

Reactivity refers to the situation where people know they are being observed so they modify their behavior. One specific type of reactivity that has been discussed and debated a lot in the behavioral sciences is the Hawthorne effect. This is a form of reactivity where subjects change their behaviors temporarily (usually in a positive way) because someone is paying attention to them. While it may not be possible to control demand and reactivity effects in all situations, the important thing is to be mindful that they occur and take steps to prevent them if possible since they can potentially impact both the validity and reliability of study results.

Validity and reliability are also related to study procedures. It may be the case that the order in which study activities are carried out changes the validity and/or reliability of the data that is collected. For instance, in the IIR Research Scenario, subjects were asked to indicate their familiarities with different search topics. *When* this question is asked is likely to impact subjects' responses. If this question is asked after subjects search, then their responses will likely be affected by their experiences searching. In a longitudinal naturalistic study, asking subjects to reflect on their searching activities at monthly intervals is unlikely to yield the same kind of data as asking them to do this at weekly intervals.

Validity and reliability can also be used to critique instruments and measures. In most cases, instruments are used to collect data that will then be used to create measures, so in some ways these things are inextricably linked. Thus, these terms will be used somewhat interchangeable in this section. Instruments that yield qualitative data are generally thought to be more valid, but less reliable than those that yield quantitative data, especially with respect to eliciting information from subjects. For instance, open-ended questions which might be used for interviewing or as part of a questionnaire do not suggest appropriate

answers and topics to subjects or force them to respond in a specific way. Instead, subjects are able to provide any information they feel is relevant and they are able to describe their attitudes and feelings in more ways than just a number. However, it is unlikely that subjects will respond to such questions in the same way at two points in time.

Instrumentation is one of the biggest threats to the internal validity of a study. Consider an example where a researcher uses a logger to record what a user does while searching, but is unaware that the logger is not really recording everything that is happening. Measures computed from the data collected via this logger will not be valid. The instrument and measures may actually be reliable; that is, they will yield consistently invalid results. Thus, it is possible to have an instrument or measure that yields reliable results, but not valid results. Reliability is a necessary (but not sufficient) condition for validity, but the converse is not true.

Instrumentation and measurement are two very big problems in IIR that need increased attention. Instrumentation and measurement are particularly tricky when studying user perceptions, attitudes and behaviors because these things can be influenced by the *process* of instrumentation and measurement (see previous discussion of method variance). In IIR, the questionnaire is one of the most widely used instruments for collecting data from subjects. It is well-known that people exhibit a number of biases when responding to questionnaire items, including social desirability responding and acquiescence. It is also well-known that people are sensitive to characteristics of measurement tools and the contexts in which they are used. Such biases are a huge source of measurement error, which poses a serious threat to the internal validity of a study. Despite this, most of the questionnaires and scales that are used in IIR do not have established validity and reliability and are often developed *ad-hoc*. While there are some items and scales that appear in many studies and could be characterized as a core question set, most of these items have not undergone any significant validity and reliability testing. They have become the core by default rather than because of their specific properties.

Theoretically, the validity and reliability of all measures should be established. This is not a trivial endeavor and requires a number of

studies designed exclusively around the measure. Practically speaking, the majority of IIR measures do not have demonstrated validity and reliability, although many have *face validity*. With respect to measurement, four major types of validity can be evaluated: face validity, predictive validity, construct validity and content validity. *Face validity* is not evaluated formally, but is related to whether the measure makes sense and is acceptable to a community of researchers. For instance, using shoe size as a measure of system usability has no face validity — it does not make sense to use such a measure as a surrogate for usability. Face validity, in many ways, is socially constructed and dependent on researcher consensus.

Predictive validity is the extent to which a measure predicts a person's behavior. For instance, if a person scores high on a college entrance exam, we would expect that person to do well in college. Thus, predictive validity looks at the relationship between the item used to measure a behavior and the behavior itself. *Construct validity* is the extent to which a measure makes sense within the context of other measures that are related to it. For instance, if a researcher develops a new item for measuring ease of use, responses to this item should be in accord with responses to other related items about ease of use. If there are five other ease of use items, responses to the new item should classify the system in a way similar to the other items. Otherwise, they are probably not all measuring the same thing. Finally, *content validity* is related to the extent to which a measure covers the range of possible meanings of the concept that it is purported to measure. Usability offers another good example to illustrate this type of validity — if we only used a single item to measure usability this measure would not have good content validity because the concept of usability is known to be complex and multi-faceted.

While validity is primarily concerned with whether or not a measure adequately captures the essence of a concept, reliability is primarily concerned with whether or not the measure yields consistent findings. Issues of reliability are complicated in situations where data is self-reported by subjects and where the researcher is the instrument. In these situations, personal bias, response bias, memory and demand effects can impact reliability. One of the best approaches to

measurement is to use instruments with established reliability. Many behavioral science disciplines have collections of established measures, some of which will be more or less appropriate to IIR depending on the focus of the study. There are some usability scales and measures that have established reliability in the human–computer interaction and business information system literatures. However, the appropriateness of these measures to IIR systems should be closely evaluated; it is likely that there are other things that need to be evaluated in the IIR situation, so the measures may actually lose some of their content validity when applied to IIR.

There are many ways to explore and establish the validity and reliability of measures, but a discussion of these techniques is beyond the scope of this paper. If a researcher is interested in establishing a new method or measure, further reading about different techniques is recommended (see [105, 286]).

13

Human Research Ethics[1]

Any research that involves human subjects necessarily requires some special ethical considerations. While ethical guidelines vary greatly from country to country and even from institution to institution, the goal of this section is to present general ethical issues that are relevant to IIR research and to stimulate more serious discussion of these issues.

Discussions of research ethics, especially in an international context, are difficult because ideas of what is right and wrong are fundamentally social constructs and can vary widely from culture to culture. Many countries do not have ethics review boards. Moreover, formalized ethics review processes are often viewed negatively by many researchers. Some view the ethics review process as mere institutional bureaucracy. Some assume that there are no ethical issues associated with IIR

[1] Much of what is presented in this section is based on my participation on the Institutional Review Board (IRB) at the University of North Carolina. Through this participation, I engage regularly in discussion, debate and reflection about the ethics of a wide-variety of research studies and procedures. It should come as no surprise that I am a strong proponent and defender of human research ethics and IRBs. This perspective will be clear in this section. It is also important to note that this paper is written from the perspective of an academic researcher working in the United States.

research since subjects are not being injected with fluids or consuming experimental medicines. In most cases, chances of physical harm to subjects in IIR evaluations are practically non-existent, but this does not mean that the ethics of IIR research should not be reviewed or that no risks exist. The principle risk to subjects in IIR evaluations is psychological harm. For example, a subject who is unable to use a retrieval system successfully may become distressed or leave the study feeling like a failure.

Providing the best possible protection to human subjects should be taken seriously since research would not be possible without them. We should be proactive with respect to evaluating the ethics of our own research and not wait for someone else to identify possible problems. As researchers, we have a responsibility to monitor our actions critically and develop ethical standards and principles for our specific research context.

13.1 Who is a Human Subject?

All of the studies depicted in Figure 2.1 with the exception of those at the systems end involve human subjects. A human subject is defined by US Federal Regulation as "a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information" (45 CFR 46.102(f)). Thus, according to the first part, all studies that involve humans — whether humans are studied directly or helping to develop research infrastructure by providing relevance assessments — are obliged to be reviewed. One might argue that a study where people provide relevance assessments is not the same as a study where humans are given experimental drugs or a study where humans' search behaviors are logged. Of course, these types of studies are not the same, but what is common is that each study cannot be done without the participation of humans. One might also argue that the point of one's IIR study is to evaluate the system and not the person, but this is missing the point — if you are using people to evaluate your system then you are necessarily studying search behavior and therefore have

an obligation to protect your participants' welfare.[2] Even the rights of TREC assessors must be protected!

13.2 Institutional Review Boards

The primary way that research involving human subjects is reviewed in the United States at academic institutions is via Institutional Review Boards (IRBs). The main purpose of IRBs is to "protect the rights and welfare of research subjects and to function as a kind of ethics committee focusing on what is right and wrong and on what is desirable or undesirable. The IRB is thus required to make judgments about what individuals and groups ought to do and how they ought to do it" [8, pp. 7–8].

Many countries have bodies similar to the IRB, especially for medical research, and there are two international documents that discuss research ethics, the Nuremberg Code[3] and the Declaration of Helsinki.[4] Although the Nuremberg Code was a response to so-called medical research by Nazi physicians, it has some important things to say about how subjects should be treated generally. Specifically, the Code articulated the requirement for voluntary and informed consent, a risk/benefit analysis that favored benefits, and the right to withdraw from the study without penalty. These three requirements form the basis of the current review process and researchers are obliged to make provisions to ensure that these things are met, regardless of the risks involved through participation. The Declaration of Helsinki, which was drafted in 1964 by the World Medical Association, extended the Nuremberg Code. This Declaration extended the Code by stating that the interests of the subject should always be given a higher priority than those of society.

In the United States, there is great variability in how different Boards operate, but there are some fundamental policies and procedures that IRBs must adhere to that are mandated by the

[2] Review happens at many different levels and research that only has humans make relevance assessments would likely be exempt from extensive review.
[3] http://www.ushmm.org/research/doctors/Nuremberg_Code.htm
[4] http://www.wma.net/e/policy/17-c_e.html

National Research Act of 1974. IRBs are formally defined by a federal regulation that describes how government agencies must operate. Since many institutions accept funding from the federal government, they are required to follow the federal regulation to institute ethics review boards to ensure the welfare and protection of human subjects. Regardless of whether a researcher accepts federal funding or not, most academic institutions require review of all research involving human subjects.

The National Research Act of 1974 was passed in response to a series of unethical events involving research with human subjects and the growing concern that human subjects were being exploited and harmed regularly in research. This response was not only motivated by events in medical research, but also research in social and behavioral sciences. The National Research Act of 1974 defined IRBs as they are currently known and established the policies and procedures that such Boards must follow when reviewing research proposals. The Act also resulted in the creation of the Belmont Report which identified three primary ethical principles that should guide the conduct of research with human subjects [228]. These principles form the basis of rules and regulations IRBs in the United States use to evaluate the ethics of research proposals.

13.3 Guiding Ethical Principles

The Belmont Report established three principles: (1) respect for persons (2) beneficence, and (3) justice. These principles are only presented briefly in this paper and it is important to note that the issues related to each principle are much more complex and nuanced than this presentation permits.

Of the three principles, *respect for persons* is the one with which IIR researchers should be most concerned. Respect for persons incorporates two ethical convictions. First, individuals should be treated as autonomous agents, and second, persons with diminished autonomy are entitled to special protection. In other words, people should be able to exercise free-will with respect to both joining and exiting a study, and people who are in positions where this free-will might be compromised

should be given special attention (e.g., those who have cognitive disabilities, children, prisoners, and those with very low education). Diminished autonomy can also be a function of the person's relationship with the researcher. For instance, if a researcher is also a teacher, students in the researcher's class are said to have diminished autonomy because of the researcher's role as teacher — the teacher has a conflict of interest, and the student may feel coerced. Asking one's students to be study participants is viewed by many as unethical. Coercion is anything that compromises a person's ability to exercise free-will and act voluntarily. This further includes offering unjustifiably large sums of money to study subjects.

Many of the issues related to respect for persons are codified in the consent form that subjects sign. The consent that is obtained from subjects is referred to as *informed consent*, which means two things: subjects are told explicitly what will happen during the study and subjects agree to this. Being informed, of course, means that subjects are provided with enough information to make a reasonable choice about whether to participate. This includes letting subjects know how many others will participate (e.g., being 1 of 60 is different from being 1 of 5). It is also important to let people know that they are free to withdraw their participation at any time without penalty and ask that their data be deleted. If compensation is provided, then it should be made clear how this will be handled. Although it is rarely the case that subjects withdraw from IIR studies, it is nevertheless important to let them know that they have this right.

The consent form should also describe to subjects how their privacy will be protected and how confidentiality will be maintained. Researchers have an obligation to protect the privacy of their subjects. A breach of such privacy has the potential to put subjects in harm's way by compromising "a person's reputation, financial status, employability, or insurability, or in some way result in stigmatization or discrimination" [8, p. 28]. It should be clear to subjects in the consent form how their data will be used and how it will be protected since a breach may have serious consequences. It is often difficult for a researcher (or subject) to predict how a breach will be harmful, so great care should be taken to protect data even when it is viewed as

relatively harmless. Things change over time, which is another thing that makes research ethics complex and continual review necessary.

Two of the best ways to protect subjects' privacy is not to associate their names with the data (i.e., do not write a subject's name on the top of a questionnaire), and as soon as the study is finished delete all records containing subjects' names and other identifiable information, such as email addresses. In most IIR studies, there is no reason to maintain records containing identifiable data. In many cases, the only record linking subjects to a specific study is the signed consent form. Because of this, many IRBs grant a *waiver of written consent* for studies that involve minimal risk to subjects. Instead of completing a written consent form, subjects are presented with what is called an *information sheet*, which looks nearly identical to a written consent form, except that it does not have a place for the subject and researcher to sign their names. Researchers can afford subjects with the most protection by not keeping any record of their participation.

Beneficence is concerned with the well-being of research subjects. At the core of this principle is the notion that subjects should not be harmed. A risk/benefit analysis of the potential risk of harm to subjects and the potential benefits of the research is usually conducted. Often, individual subjects do not benefit directly from participating in a study (note that receiving compensation is not a benefit) and benefits are often discussed at a societal level. It is uncommon in IIR evaluations for individual subjects to benefit directly from their participation, although it is conceivable that subjects might learn how to be better searchers. Societal benefits — i.e., helping to create better information access systems — are usually the biggest benefit of IIR evaluations. Although most IIR studies do not involve great risk of harm, it is important to consider the psychological risks associated with participating in a study, as well as the risks involved with a breach of privacy once the data have been collected.

Another issue related to beneficence is the quality of a research project: if a project is so poorly designed that it results in invalid and unreliable data, is this an ethical concern? Although the answer to this question is debatable, it can be viewed as an ethical concern since subjects have been put at some risk (however minimal) through their

participation with no possibility of benefit (even to society) and, more importantly, their time has been wasted. Of course, all studies have flaws. Many studies fail to support research hypotheses, so "waste" both subjects' and researchers' time *inadvertently*. The issue is whether the researcher has done everything to ensure that the study design is the best possible before commencing. When little or no effort or thought has been put into planning and testing a study method, then the question of whether the researcher has violated ethical obligations to subjects can be asked.

The final principle is *justice*, which is related to who bears the burden of the research and who is able to benefit from it. This principle addresses the practice of targeting and recruiting vulnerable subjects, such as poor, uneducated people or people with diminished autonomy, such as prisoners. Although this principle is usually not an issue in IIR studies, we can consider the implications of only studying a particular group of subjects — for instance, university students. While university students arguably bear the heaviest burden in IIR studies (especially for those conducted in academic settings), they are also likely to benefit from this research in the long-term. However, other, less frequently studied groups might also be able to contribute to what we know about IIR. If these users are not studied, then we may miss the needs of these groups and eventually fail to incorporate these into the IIR systems and techniques we develop.

13.4 Some Specific Concerns for IIR Researchers

In addition to some of the general concerns for researchers presented in the preceding section, there are special issues related to the IIR research context that deserve mention. These are issues which will likely be discussed in more detail in the upcoming years. The first issue is related to *user privacy and search logs* [136]. Each type of logging situation presents a different set of concerns. Although many readers will automatically think of the large search logs amassed by search engine companies, logging is a part of most IIR studies and logs contain varying amounts of personal information, which when released might cause harm to users. The logs generated by users who are evaluating

an experimental IR system on a closed corpus with assigned tasks are probably the least risky. As the nature of the study changes from closed corpora and assigned tasks to the open Web and natural tasks, the risk of harm and privacy violations increases and the researcher has a greater responsibility to protect the data and subjects.

In 2006, AOL released search log data to the public and it was immediately clear that individual users could be identified through deductive disclosure [118]. The people at AOL probably had good intentions — they were, after all, releasing a large data set that could be used by researchers — but this case highlighted the sensitivity of search log data and how relatively easy it is to reconstruct individual identities by putting together lots of smaller pieces of information. The problem of deductive disclosure is challenging and was addressed in part by a workshop on query log analysis at the *World Wide Web Conference* in 2007 [10]. Example research from this workshop include a discussion of query log anonymization and solutions [2], and an analysis of query logs with a focus on privacy and the applicability of using existing privacy guidelines such as HIPAA[5] [298]. There is a desire to share collections and shared collections have traditionally been an important part of IR research. However, the de-identification process is much more complicated with personal data and it is likely that more work will need to be done on the anonymization process before such data can be made publicly available.

There are other ethical issues related to logging user interactions that are less obvious. First, consider the principle of informed consent — when users accept cookies or agree to conditions of use for search engine plug-ins, are they really giving *informed* consent [196]? Are such agreements clear about how much privacy the user gives up when using particular Web sites or search engines? Furthermore, while users typically grant permission for researchers to log their interactions, there are many interactions that involve more than one person and third-party disclosure becomes an issue. For instance, a researcher who studies retrieval of email messages may obtain permission from a set of users to log their email, but it is unlikely that the researcher will

[5] http://www.hhs.gov/ocr/hipaa/

obtain permission from all people sending that user email. A similar thing can be said for the use of personal photo and video collections — such items typically contain images of a lot of people. There are also other issues with regard to what people search for and look at online. Researchers who do survey research are required to provide assistance to subjects if they respond in certain ways to certain questions, for example, questions assessing suicidal ideation. What if a subject looks at documents about suicide or bomb-making? Do we have any special ethical obligations to the subject or to society?

Another set of ethical issues stems from crawling and using postings, comments, messages, reviews, interactions, etc. that have been created by users — for instance, as part of myspace.com, amazon.com, and in other environments that promote social interaction. Some of the first researchers to study internet groups and online communication were from the communication and sociology fields [85, 98, 214, 252, 289]). These researchers discovered that collecting and analyzing data that people post online in "public" venues was not as straightforward as it seemed; many people who posted messages became upset with how their messages were used and repurposed. Other controversial methods involved researchers joining groups as legitimate participants only to gather data for research. Although it is likely to change in the future, current review boards first question whether the site or service has any policy that forbids research-related activities, and then looks at the extent to which users have to authenticate to participate. There is often a distinction made between public and private when authentication is required because it can potentially change users' expectations about how their information will be used and who will consume it.

Another important issue that has not been discussed much in IIR is the life and death of data. Historically in IIR, data sets have been retained infinitude because collecting data from users has always been very time-consuming. Indeed, one of the benefits of conducting an IIR study is the collection of a data set that can be used for future investigations. One implication of keeping data sets for perpetuity is that data sets will outlive the researchers who assemble them. Most researchers do not have a specific plan for what will happen to their data sets over the course of their lifetime and subjects do not always have clear

expectations about how long their data will exist and how this may affect them in the future. Ultimately, the researcher is responsible for protecting the data throughout its lifetime (even if this exceeds the lifetime of the researcher) and should articulate in writing a life-plan for the data. With the emerging institutional repository movement at universities, researchers should also think of the implications of turning-over data sets to be stored in such repositories, especially with respect to ownership. If a researcher leaves a particular institution, the data set might have to be left behind. The emergence of institutional repositories and cyber-infrastructure will likely change how proposals are reviewed, how researchers communicate with subjects and how data are stored and protected.

Videotaping subjects also presents special ethical issues [189]. In the past, traditional video recorders were used to capture computer screens while subjects engaged in IIR. Typically such cameras were placed behind the subject and the back of the subject's head was usually visible on the recording. In some cases, the side of the subject's face was also visible. This necessarily changes the risks because the videos contain the likeness of the individual and as long as the videos exist, subjects will be identifiable. Most review boards require researchers to get additional permission from subjects to videotape, which allows subjects to opt-out of the videotaping. Today, most screen recording software allows researchers to record the screen of the computer without recording the subject's likeness. However, video recordings are still in use in many studies for other types of observations. Special considerations also need to be made with respect to audio recordings which are often made during interviews since a person's voice is recorded and it could be used to identify the person. Thus, if audio recordings are captured transcriptions should be made as soon as possible and the tapes destroyed to provide the most protection.

There have been a number of IIR studies that involve *deception* — subjects are lead to believe one thing about what is going on in the study, but in reality another thing is happening. These studies are not usually traditional IIR evaluations, but experimental studies that are more focused on behavior. Examples of deception include manipulating the order of search results, telling subjects that they are using a

particular system when they are not and giving false feedback. Most of the deception involved in IIR studies can be considered minor. Regardless, study proposals that involve deception are looked at more carefully by IRBs because they involve more risks. Moreover, the *informed* part of the consent process is compromised since researchers are unable to obtain consent regarding deception. One essential element in studies that involve deception is a *debriefing* at the end of the study to let subjects know the real purpose of the study and to describe the deception that was involved. Subjects should always be given an opportunity to ask questions at the end of the study and to withdraw their data if they wish.

A final ethical concern for IIR researchers is the extent to which ethics should be considered when reviewing manuscripts for publication. In some disciplines, it is common for researchers to include in published manuscripts the IRB approval number or a statement about how the ethics of the study were evaluated. In IIR, it is taken in good faith that researchers abide by ethical principles and requirements, but where, when and how are such ethical principles communicated and acquired? What guidelines are taken to be the standard? How are young researchers educated about ethics? Ethical responsibilities in general have not extended beyond local review of individual studies. However, it may be the case that ethics review will need to happen at different levels of the research process. Consider research that has been published using the data set released (and retracted) by AOL. Is it ethical to use this data? Should our community publish reports that use this data? If the answer is no, then how should we monitor this?

14

Outstanding Challenges and Future Directions

This paper provided background information about IIR and guidance about how to design and conduct IIR evaluations. However, there were many topics that could not be discussed directly and there are also many outstanding challenges.

14.1 Other Types of Systems

This paper focused on traditional text retrieval systems, although there are currently a great number of systems that support retrieval of different types of information objects including images, video, audio, and personal information; varying types of textual units such as answers and passages; varying languages; and varying devices. There are also a number of systems that use experimental visualization techniques and offer support for a broader range of information-seeking activities (e.g., saving and sorting results). While many of the basic techniques presented in this paper can be applied to different types of IIR systems and use scenarios, each have their own special set of concerns that must be addressed since the nature of the objects being retrieved and the information needs and purposes behind search vary.

Multimedia search includes image, video and audio. One interaction that is very different in these types of systems from traditional text retrieval systems is querying. While in text retrieval systems the objects that are being retrieved match the method of querying, in multimedia systems querying is still often limited to text which does not necessarily match the form of the objects being retrieved. Furthermore, notions of aboutness and relevance can be even more problematic in these settings than in traditional text-based settings. Although not technically a part of multimedia search, interfaces that use experimental visualization techniques also require special consideration [156].

Personal information management (PIM) focuses on a variety of media types, including self-created objects and email, and a variety of specialized tasks such a re-finding and information organization [86, 155]. PIM is an ongoing activity often done in anticipation of future actions (such as re-finding information objects) or expected uses (such as sharing information objects). Because PIM is concerned with information classification and retrieval, it has many things in common with IIR. However, two important differences are that a variety of types of information objects and systems might be examined in a single study and that the information is personal. The implications of this are that evaluations often have to be more flexible and tailored to individuals and carried out in naturalistic settings. This makes it more difficult to study causal relationships and to identify findings that generalize.

There are also many sub-areas of IIR that investigate how users search in environments where smaller units of information are retrieved, such as XML fragments, answers, passages and summaries, and where larger units of information are retrieved such as books. Interactive retrieval of documents marked with XML has received considerable attention in the past few years, most notably through the INEX workshop series (e.g., [176, 273]). In these studies, researchers must accommodate search of parts of documents rather than the whole and understand how users make sense of these various fragments. Research investigating interactive question–answering is also emerging as an area that has many interesting opportunities for IIR researchers. Some studies have been done to investigate users' preferences for answer sizes

[179] and to develop evaluation methods and measure for interactive QA systems [127, 164, 169], but in general, less is known about how people interact with and use QA systems. Many QA systems use natural language dialogue to facilitate interaction, which adds another dimension to the evaluation. Summarization technology is developing rapidly and there is no reason to expect that interactive, user-centered evaluations will not be of interest to researchers in this community. Finally, the success of a recent SIGIR workshop on book search [160] has demonstrated a renewed interest in a domain that is rooted in interactive IR evaluation [83].

Another area where there is quite a lot of systems-centered research, but not much user-centered research is cross-language retrieval, although several researchers have made contributions to this area [204, 211]. Cross-language retrieval is not as widespread and common as the other types of retrieval discussed in the preceding paragraphs, so it is hard to identify search tasks and contexts. However, cross-language retrieval is an important and relevant task to many, most notably government intelligence officers. In addition, there are research programs whose goal is to bring together numerous technologies, including summarization, multimedia and cross-language into a single system and some preliminary reports of user-centered evaluations of these systems [300].

This paper did not address evaluation issues associated with adaptive systems and other systems designed to personalize interactions. These types of systems are particularly difficult to evaluate because usually they are designed to be used over long periods of time. A single search session of the kind that typically happens in a standard IIR study simply does not allow such systems to realize their potential. Since search is personalized to individuals, it is also difficult to set-up a general evaluation framework for all subjects. Social search systems are also showing great promise, but introduce additional considerations. In particular, the cold-start and data sparsity problems must be addressed before evaluation can take place. However, once these problems are addressed, many aspects of the standard IIR evaluation model can be applied. With respect to experimental studies of search behaviors, social search creates many opportunities for researchers to test and

apply theories from social psychology to better understand behavior in this context.

Collaborative IR systems that support group information-seeking and retrieval have emerged recently as a popular area in IR (see [102] for an overview; [153, 199]). Researchers in computer supportive cooperative work (CSCW) and educational technologies have studied systems that support collaborative work for some time. While the research from these areas can provide guidance on the design of studies for collaborative IR, there are also a number of issues specific to the IR situation that will need to be addressed. Again, some elements of the standard IIR evaluation model might be effective in this context, but the danger is always that an overuse of such models prevents the development of more appropriate models. There is also an additional type of interaction that must be accounted for — the interaction between the people engaged in collaboration. The future will likely involve not only the development of novel systems for collaborative IR, but also novel evaluation methods and measures, which might be rooted in communication theory and social psychology.

Finally, the evaluation and study of mobile information-seeking and retrieval also introduces its own special issues [109]. The information-seeking needs and behaviors of users, the situations in which searching takes place and the nature of the device and hardware make this type of retrieval different from standard, non-mobile text search.

14.2 Collections

Sharable collections have played an important role in IR and IIR evaluation, but most of the collections that have been used in IIR studies test their limits in terms of generalizability and usability. TREC collections have been used widely in IIR evaluations, but as described earlier, researchers must make some simplifying assumptions about the nature of relevance, the generalizability of relevance assessments and appropriateness of assigned search tasks.

There are several possible directions that IIR research can take with respect to developing shared collections. The first is to determine how collections developed for systems-centered evaluation can be

better used in IIR evaluations. This involves engaging with a number of perennial problems in IIR, including the nature of relevance. The second direction is to create new collections that contain some elements of traditional collections, such as a corpus of documents, but that also contain new elements that are specific to the interactive retrieval situation. Voorhees [287] discusses the difficulties of creating a test collection for adaptive information retrieval.

The third direction is to develop task sets that can be used in different situations. While researchers often use TREC topics as search tasks, a larger variety of tasks that systematically vary across a number of attributes (e.g., difficulty and specificity) would greatly facilitate evaluation and experimentation. The development of shared tasks is more than just penning them. Like any instrument, tasks should undergo a number of tests to ensure that they are representative of the attributes they are purported to embody, that they can be used consistently across a number of situations and that users interpret the tasks in expected ways.

A final direction towards shared collections is shared data sets, which may or may not conform to the traditional definition of a collection. This includes large scale query log data collected by search engines and other large organizations, as well as data collected by researchers who focus on smaller-scale laboratory experiments. Search engine companies in particular, have made some efforts to share log data. More successful examples involve controlled sharing through personal relationships, competitive grant programs for academic researchers and more recently, specialized workshops (e.g., *Workshop on Web Search Click Data*[1]). Sharing with fewer restrictions and across more circumstances may be on the horizon, but privacy issues will likely dictate that some restrictions will always apply.

Many academic researchers have collected detailed log data, often supplemented with self-report and interview data, from subjects which can also act as a type of shared collection. These data sets typically involve few subjects, but contain rich contextual data about needs and behaviors. Some of these data sets are collected in the context of Web

[1] http://research.microsoft.com/∼nickcr/wscd09/

search, while others are collected as part of evaluations of experimental systems. Although there is not a strong tradition or incentive to share such data sets[2] and no real infrastructure to support sharing, a data repository would support numerous kinds of research, including classic IR and IIR, as well as meta-analysis, systematic review, and comparative and historical analysis.

14.3 Measures

One of the most significant measurement challenges is developing performance measures that can be used in interactive search scenarios. There are a number of standard evaluation measures available to those conducting systems-centered IR studies. Many of these have been used in IIR evaluations, but in most cases, the assumptions underlying the measures do not match what happens during interactive searching. Most of these measures assume stable, independent, binary relevance. In interactive search situations, relevance assessments can change throughout the search session and vary based on the presentation order of documents, as well as other contextual factors such as the user's familiarity with the search topic. In interactive situations, relevance assessments are rarely binary and are made by many users (not just a single assessor) who do not always agree on what is relevant. Furthermore, many standard IR evaluation measures are based on a one-to-one correspondence between a query and topic, and only accommodate a single search result list for any given topic. In interactive search scenarios, users typically enter many queries during a search session and view a number of search results lists. This situation introduces a variety of issues including duplicate search results. The work on discounted cumulated gain [149, 150] represents an important step towards the development of performance measures that are better suited to interactive searching, but more measures are needed, especially those that reflect session-based performance.

[2] The American Psychology Association [9, p. 354] includes as an ethical principle of scientific publishing that researchers maintain their data sets for at least five years after publication and make them available to journal publishers and other researchers who might question the findings and/or want to replicate the study.

Better methods for eliciting evaluative data from subjects are also needed. One approach to obtaining better evaluative data from subjects is to identify indirect measures or subsidiary measures that are highly correlated with measures of interest. An example of this is Czerwinski et al.'s [66] subjective duration assessment which was described earlier. Indirect measures are a potentially useful way to address some of the problems with self-report measures, even though many actually rely on self-report. For instance, Czerwinski et al. [66] asked subjects to estimate how long it took them to complete tasks (self-report) and used the differences between these estimates and the real time to determine task difficulty. The underlying assumption is that if subjects were asked to directly respond to a question about task difficulty, that these responses would likely be subject to response bias. It is important to note that the establishment of indirect measures requires careful and thorough investigation. Such measures must be evaluated rigorously against some gold standard, which is often challenging to elicit and/or determine. For instance, to link the discrepancy between a user's time estimate and the actual time with task difficulty requires some baseline measure of task difficulty. The use of the dual monitoring task by Dennis et al. [72] also represents an attempt to use an indirect method for understanding more about subjects' experiences.

Eye-tracking data provides another source of information that can be used to create evaluation measures. Researchers have studied eye movements for well over 100 years — initially in the fields of cognitive psychology and physiology and later in the field of human–computer interaction [143]. Jacob and Karn [143] attribute the first use of eye-tracking in human–computer interaction to Fitts et al.'s [96] study of the movements of airplane pilots' eyes as they used cockpit controls; eye movements were recorded using motion picture cameras. Today, there is better equipment and better theoretical frameworks for understanding eye movements, although there have only been a few studies that have used eye-tracking in IIR research (e.g., [152, 222]). Although the equipment is clearly better today than in the past, it is still expensive and awkward for subjects. Even with the best equipment, subjects often need to sit very still, which can be difficult when searching for multiple tasks during a one hour period. Another difficulty is that large

amounts of data are generated — it can be difficult to analyze and make-sense of this data. However, eye-trackers provide more refined information about how a subject experiences an IIR system and conducts information-seeking and retrieval. This includes more detailed information about which parts of documents subjects view and if subjects cognitively engage with a particular feature or object even when there is no observable log action such as a click. Lorigo et al. [184] provide an overview of eye-tracking and online search and identifies future research directions.

Emotional and affective measures are likely to play an increasingly important role in IIR evaluation. The notion of affective computing has been around for quite some time in the human–computer interaction literature [212], but has not made its way to IIR research [157]. There are certainly studies of users' emotions and affective states during the information-seeking process (e.g., [174]), but researchers have yet to tie specific emotional responses to particular IR interactions and states. Arapakis and Jose [12] recently conducted a study which documented the range of emotions that subjects experience while engaged in laboratory IR tasks. While the ultimate goal of this work is to use emotions as feedback for retrieval, the work suggests that emotional or affective measures might be useful for evaluative purposes. What remains is for someone to develop a theoretical framework for understanding how emotions and affective responses can be used as evaluative feedback and how one might reliably capture such information, whether through facial recognition software or self-report. In addition to measures of affect and emotion, Hassenzahl et al. [124] discuss the notion of hedonic quality. Hedonic qualities are qualities such as originality, aesthetic appeal and innovativeness that do not necessarily have any relation to the task the system is designed to support or the system's performance, but that still contribute to people's experiences and evaluations.

Physiological signals, such as heart rate and perspiration are also potentially rich sources of data about users' experiences and reactions during IIR. Researchers in many disciplines have investigated the relationship between human physiological signals and emotional and mental states. In IIR, such signals might be used as evaluation measures or implicit relevance feedback. Equipment for measuring basic signals such

as skin response and heart rate are relatively inexpensive. Although such equipment is not a normal part of most users' workspaces, physiological sensors are increasingly available and it is not difficult to imagine a world where these types of sensors are a normal part of people's lives. As with eye-tracking data, the biggest challenge with physiological data is analyzing the large number of signals and understanding what they really mean.

15

Conclusion

Reflecting on three decades of IR research, Robertson [220, p. 447] notes, "In the end, any experimental design is a compromise, a matter of balancing those aspects which the experimenter feels should be realistic with those controls which are felt to be necessary for a good result". Similarly, research design in IIR is about making choices; the primary goal of this paper was to catalog and compile material related to methods for the evaluation of interactive information retrieval systems into a single source to help researchers make more informed design decisions. Robertson [220, p. 447] continues, "a field advances not by deciding on a single best compromise, but through different researchers taking different decisions, and the resulting dialectic". The intent of this paper is not to suggest that there is a single best evaluation method or even that evaluation is the only useful type of IIR research — IIR is more than system evaluation and retrieval effectiveness. IIR requires pluralistic approaches and methods. A single, prescribed model would be deleterious.

Despite the length of this paper, many of the presentations were brief; it is hoped that this paper will provide a foundation around which others can discuss methods for studying IIR. This includes the creation

of more detailed reviews of some of the topics discussed in this paper such as IIR history, measures and ethics. People have varying opinions about how IIR evaluation should be conducted. The content of this paper represents one such opinion that is informed heavily by the literature, the author's research experiences and an academic background that is rooted in the behavioral sciences. IIR blends behavioral and computer sciences in an effort to study very complex activities: information search and retrieval. It can be difficult to negotiate these two research traditions and uphold their respective research standards all while maintaining scientific integrity. The length of this paper reflects the complexity, difficulties and nuances of studying IIR and demonstrates why more serious scholarship devoted specifically to methods and measures is needed to further IIR research.

Acknowledgments

I would like to thank Nick Belkin and Paul Kantor for their training and guidance; Justin Zobel, Barbara Wildemuth and Cassidy Sugimoto for their feedback and discussion about this paper; Fabrizio Sebastiani and Jamie Callan for their great patience and encouragement; and three anonymous reviewers for their careful and thoughtful comments.

References

[1] R. P. Abelson, *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Publishers, 1995.

[2] E. Adar, "User 4XXXXX9: Anonymizing query logs," in *Proceedings of the Workshop on Query Log Analysis: Social and Technological Challenges, at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[3] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior," in *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 3–10, Seattle, WA, 2006.

[4] J. Allan, "HARD track overview in TREC 2003: High accuracy retrieval from documents," in *TREC2003, Proceedings of the 12th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington, DC: GPO, 2004.

[5] J. Allan, "HARD track overview in TREC 2005: High accuracy retrieval from documents," in *TREC2005, Proceedings of the 14th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington, DC: GPO, 2006.

[6] B. Allen, "Information needs: A person-in-situation approach," in *Information Seeking in Context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, (P. Vakkari, R. Savolainen, and B. Dervin, eds.), pp. 111–122, Tampere, Finland, 1997.

[7] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, pp. 10–16, 2008.

[8] R. Amdur, *Institutional Review Board Member Handbook*. Sudbury, Massachusetts: Jones and Bartlett Publishers, 2003.

[9] American Psychological Association, *Publication Manual of the American Psychological Association*. Washington, DC: APA, Fifth ed., 2001.

[10] E. Amitay, G. C. Murray, and J. Teevan, "Workshop on query log analysis: Social and technological challenges," in *Proceedings of the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[11] P. Anick, "Using terminological feedback for web search refinement: A log based study," in *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 88–95, Toronto, CA, 2003.

[12] I. Arapakis and J. Jose, "Affective feedback: An investigation of the role of emotions during an information seeking process," in *Proceedings of the 31st Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 395–402, Singapore, Malaysia, 2008.

[13] E. Babbie, *The Practice of Social Research*. CA, Wadsworth, 10 ed., 2004.

[14] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, "Relevance assessment: Are judges exchangeable and does it matter," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 667–674, Singapore, Malaysia, 2008.

[15] M. J. Bates, "Information search tactics," *Journal of the American Society for Information Science*, vol. 30, pp. 205–214, 1979.

[16] M. M. Beaulieu, "Interaction in information searching and retrieval," *Journal of Documentation*, vol. 56, pp. 431–439, 2000.

[17] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams, "Okapi at TREC-5," in *Proceedings of the 5th Text Retrieval Conference* (*TREC-5*), (E. M. Voorhees and D. K. Harman, eds.), pp. 143–165, Washington, DC: GPO, 1997.

[18] B. Bederson, "Interfaces for staying in the flow," *Ubiquity*, vol. 5, 2004.

[19] N. J. Belkin, "Anomalous states of knowledge as a basis for information retrieval," *Canadian Journal of Information Science*, vol. 5, pp. 133–143, 1980.

[20] N. J. Belkin, "Helping people find what they don't know," *Communications of the ACM*, vol. 43, pp. 58–61, 2000.

[21] N. J. Belkin, A. Cabezas, C. Cool, K. Kim, K. B. Ng, S. Park, R. Pressman, S. Rieh, P. Savage, and I. Xie, "Rutgers interactive track at TREC-5," M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams, "Okapi at TREC-5," in *Proceedings of the 5th Text Retrieval Conference* (*TREC-5*), (E. M. Voorhees and D. K. Harman, eds.), pp. 257–265, Washington, DC: GPO, 1997, 1997.

[22] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool, "Query length in interactive information retrieval," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 205–212, Toronto, Canada, 2003.

[23] N. J. Belkin and A. Vickery, *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-Based Systems*. Library and Information Research Report 35: The British Library, University Press, Cambridge, 1985.

[24] D. J. Bell and I. Ruthven, "Searchers' assessments of task complexity for Web searching," in *Proceedings of the 26th Annual International European Conference on Information Retrieval* (*ECIR 2004*), pp. 57–71, Sunderland, UK, 2004.

[25] J. L. Bennett, "The user interface in interactive systems," *Annual Review of Information Science and Technology*, vol. 7, pp. 159–196, 1972.

[26] J. D. Bernal, "Preliminary analysis of pilot questionnaires on the use of scientific literature," *The Royal Society Scientific Information Conference*, pp. 589–637, 1948.

[27] W. H. Beyer, *Handbook of Tables for Probability and Statistics*. Cleveland, OH: Chemical Rubber Co. Publishers, 1966.

[28] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds: Identifying relevant websites from user activity," in *Proceedings of the 17th International Conference on the World Wide Web* (*WWW '08*), pp. 51–60, Beijing, China, 2008.

[29] A. Blandford, A. Adams, S. Attfield, G. Buchanan, J. Gow, S. Makri, J. Rimmer, and C. Warwick, "The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work," *Information Processing and Management*, vol. 44, pp. 4–21, 2008.

[30] C. L. Borgman, "End user behavior on an online information retrieval system: A computer monitoring study," in *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '83*), pp. 162–176, Bethesda, MD, 1983.

[31] C. L. Borgman, "All users of information retrieval systems are not created equal: An exploration into individual differences," *Information Processing and Management*, vol. 25, pp. 237–251, 1989.

[32] P. Borlund, "Experimental components for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 56, pp. 71–90, 2000.

[33] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science*, vol. 54, pp. 913–925, 2003a.

[34] P. Borlund, "The IIR evaluation model: A framework for evaluation of interactive information retrieval systems," *Information Research*, vol. 8, p. 152, 2003b.

[35] P. Borlund and P. Ingwersen, "The development of a method for evaluating interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[36] P. Borlund and P. Ingwersen, "Measure of relative relevance and ranked half-life: Performance indicators for interactive information retrieval," in *Proceedings of the 21st ACM SIGIR Conference on Research and Development of Information Retrieval* (*SIGIR '98*), pp. 324–331, Melbourne, Australia, 1998.

[37] P. Borlund and I. Ruthven, "Introduction to the special issue on evaluating interactive information retrieval systems," *Information Processing and Management*, vol. 44, pp. 1–3, 2008.

[38] P. J. Borlund, W. Schneider, M. Lalmas, A. Tombros, J. Feather, D. Kelly, A. P. de Vries, and L. Azzopardi, *Proceedings of the 2nd International Symposium on Information Interaction in Context*. London, UK, 2008.

[39] B. R. Boyce, C. T. Meadow, and D. H. Kraft, *Measurement in Information Science*. London, UK: Academic Press, Inc, 1994.

[40] J. Bradley, "Methodological issues and practices in qualitative research," *Library Quarterly*, vol. 63, pp. 431–449, 1993.

[41] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," in *Proceedings of the 5th International Conference on Intelligent User Interfaces* (*IUI '00*), pp. 44–51, New Orleans, LA, 2000.

[42] M. Bulmer, *Questionnaires V. 1.* Thousand Oaks, CA: Sage Publications, 2004.

[43] K. Byström, "Information and information sources in tasks of varying complexity," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 581–591, 2002.

[44] K. Byström and P. Hansen, "Conceptual framework for tasks in information studies," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 1050–1061, 2005.

[45] K. Byström and K. Järvelin, "Task complexity affects information seeking and use," *Information Processing and Management*, vol. 31, pp. 191–213, 1995.

[46] J. Callan, J. Allan, J. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai, "Meeting of the MINDS: An information retrieval research agenda," *SIGIR Forum*, vol. 41, pp. 25–34, 2007.

[47] D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

[48] R. Capra, "Studying elapsed time and task factors in re-finding electronic information," *Personal Information Management, CHI 2008 Workshop*, Florence, Italy, 2008.

[49] D. O. Case, *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. Lexington, KY: Academic Press, 2002.

[50] K. Charmaz, "Qualitative interviewing and grounded theory analysis," in *Handbook of Interview Research: Context and Method*, (J. F. Gubrium and J. A. Holstein, eds.), CA: Sage Publications, 2002.

[51] E. Chatman, "The impoverished life-world of outsiders," *Journal of the American Society for Information Science*, vol. 47, pp. 193–206, 1996.

[52] H. Chen, R. Wigand, and M. Nilan, "Exploring Web users' optimal flow experiences," *Information Technology and People*, vol. 13, pp. 263–281, 2000.

[53] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human–computer interface," in *Proceedings of ACM Human Factors in Computing Systems Conference* (*CHI 1988*), pp. 213–218, 1988.

[54] C. W. Cleverdon, "The Cranfield tests on index language devices," in *Readings in Information Retrieval*, (K. Spark-Jones and P. Willett, eds.), (Reprinted from *Aslib Proceedings*, pp. 173–192.) San Francisco: Morgan Kaufman Publishers, 1997/1967.

[55] C. W. Cleverdon, L. Mills, and M. Keen, *Factors Determining the Performance of Indexing Systems, vol. 1 — Design*. Cranfield, England: Aslib Cranfield Research Project, 1966.

[56] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, Second ed., 1988.

[57] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Quarterly*, vol. 19, pp. 189–211, 1995.

[58] C. Cool, "The concept of situation in information science," *Annual Review of Information Science and Technology*, pp. 5–42, 2001.

[59] W. S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems," *American Documentation*, vol. 19, pp. 30–41, 1968.

[60] W. S. Cooper, "On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation," *Journal of the American Society for Information Science*, vol. 24, pp. 87–100, 1973.

[61] P. Cowley, J. Haack, R. Littlefield, and E. Hampson, "Glass Box: Capturing, archiving and retrieving workstation activities," in *Proceedings of the 2nd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences* (*CARPE '05*), pp. 13–18, Santa Barbara, CA, 2006.

[62] F. Crestani and H. Du, "Written versus spoken queries: A qualitative and quantitative comparative analysis," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 881–890, 2006.

[63] M. Csikszentmihalyi, *Finding Flow: The Psychology of Engagement with Everyday Life*. New York: Basic Books, 1997.

[64] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," *Journal of Nervous and Mental Disease*, vol. 175, pp. 526–536, 1987.

[65] M. Czerwinski, E. Horvitz, and E. Cutrell, "Subjective duration assessment: An implicit probe for software usability," in *Proceedings of IHM-HCI 2001 Conference*, pp. 167–170, Lille, France, 2001.

[66] M. Czerwinski, E. Horvitz, and S. Wilhite, "A diary study of task switching and interruptions," in *Proceedings of ACM Human Factors in Computing Systems Conference* (*CHI '04*), pp. 175–182, Vienna, Austria, 2004.

[67] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies: Why and how," in *Proceedings for the 1st International Conference on Intelligent User Interfaces* (*IUI '93*), pp. 193–200, Orlando, FL, 1993.

[68] H. Dang, D. Kelly, and J. Lin, "Overview of the TREC 2007 question answering track," in *TREC2007, Proceedings of the 16th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington DC: GPO, 2008.

[69] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 319–340, 1989.

[70] M. De Mey, "The cognitive viewpoint: Its development and its scope," in *CC77: International Workshop on the Cognitive Viewpoint,* (M. De Mey et al., eds.), pp. xvi–xxxi, Ghent, Belgium: University of Ghent Press, 1977.

[71] S. Debowski, R. Wood, and A. Bandura, "The impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic enquiry," *Journal of Applied Psychology*, vol. 86, pp. 1129–1141, 2001.

[72] S. Dennis, P. Bruza, and R. McArthur, "Web searching: A process-oriented experimental study of three interactive search paradigms," *Journal of the*

American Society for Information Science and Technology, vol. 53, pp. 120–133, 2002.

[73] N. K. Denzin and Y. S. Lincoln, *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications, 2000.

[74] B. Dervin, "From the mind's eye of the user: The sense-making qualitative-quantitative methodology," in *Qualitative Research in Information Management*, (R. Glazier, ed.), pp. 61–84, Englewood, CO: Libraries Unlimited, 1992.

[75] B. Dervin, "Given a context by any other name: Methodological tools for taming the unruly beast," in *Information Seeking in Context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, pp. 13–38, Tampere, Finland, 1996.

[76] A. Dillon, "User analysis in HCI: The historical lesson from individual differences research," *International Journal of Human–Computer Studies*, vol. 45, pp. 619–637, 1996.

[77] W. D. Dominick and W. D. Penniman, "Automated monitoring to support the analysis and evaluation of information systems," in *Proceedings of the 2nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '79*), pp. 2–9, Dallas, TX, 1979.

[78] P. Dourish, "What we talk about when we talk about context," *Personal and Ubiquitous Computing*, vol. 8, pp. 19–30, 2004.

[79] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've Seen: A system for personal information retrieval and re-use," in *Proceedings of the ACM Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 72–79, Toronto, Canada, 2003.

[80] S. T. Dumais and N. J. Belkin, "The TREC interactive tracks: Putting the user into search," in *TREC: Experiment and Evaluation in Information Retrieval*, (E. M. Voorhees and D. K. Harman, eds.), pp. 123–153, Cambridge, MA: MIT Press, 2005.

[81] M. Dunlop, "Time, relevance and interaction modeling for information retrieval," in *Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval* (*SIGIR '97*), pp. 206–213, Philadelphia, PA, 1997.

[82] M. D. Dunlop, C. W. Johnson, and J. Reid, "Exploring the layers of information retrieval evaluation," *Interacting with Computers*, vol. 10, pp. 225–236, 1998.

[83] D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum, "Formative design-evaluation of SuperBook," *ACM Transactions on Information Systems*, vol. 7, pp. 30–57, 1989.

[84] M. Eisenberg, "Measuring relevance judgments," *Information Processing and Management*, vol. 24, pp. 373–389, 1988.

[85] D. Elgesem, "What is special about the ethical issues in online research?," *Ethics and Information Technology*, vol. 4, pp. 195–203, 2002.

[86] D. Elsweiler and I. Ruthven, "Towards task-based personal information management evaluations," in *Proceedings of the 30th ACM Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 22–30, Amsterdam, The Netherlands, 2007.

[87] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press, Revised ed., 1993.

[88] C. H. Fenichel, "Online searching: Measures that discriminate among users with different types of experience," *Journal of the American Society for Information Science*, vol. 32, pp. 23–32, 1981.

[89] K. D. Fenstermacher and M. Ginsburg, "Client-side monitoring for Web mining," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 625–637, 2003.

[90] R. Fidel, "Online searching styles: A case-study-based model of searching behavior," *Journal of the American Society for Information Science*, vol. 35, pp. 211–221, 1984.

[91] R. Fidel, "Qualitative methods in information retrieval research," *Library and Information Science Research*, vol. 15, pp. 219–247, 1993.

[92] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," *Transactions on Information Systems*, vol. 20, pp. 116–131, 2002.

[93] K. E. Fisher, S. Erdelez, and L. E. F. McKenchnie, *Theories of Information Behavior*. Medford, NJ: Information Today, 2005.

[94] D. W. Fiske, "Convergent-discriminant validation in measurements and research strategies," in *Forms of Validity in Research*, (D. Brinbirg and L. H. Kidder, eds.), pp. 77–92, San Francisco: Jossey-Bass, 1982.

[95] S. T. Fiske, "Mind the gap: In praise of informal sources of formal theory," *Personality and Social Psychology Review*, vol. 8, pp. 132–137, 2004.

[96] P. M. Fitts, R. E. Jones, and J. L. Milton, "Eye movements of aircraft pilots during instrument-landing approaches," *Aeronautical Engineering Review*, vol. 9, pp. 24–29, 1950.

[97] B. N. Flagg, *Formative Evaluation for Educational Technologies*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

[98] S. Flicker, D. Haans, and H. Skinner, "Ethical dilemmas in research on Internet communities," *Qualitative Health Research*, vol. 14, pp. 124–134, 2004.

[99] N. Ford, D. Miller, and N. Moss, "The role of individual differences in Internet searching: An empirical study," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 1049–1066, 2001.

[100] N. Ford, D. Miller, and N. Moss, "Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes and approaches," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 741–756, 2005.

[101] N. Ford, T. D. Wilson, A. Foster, D. Ellis, and A. Spink, "Information seeking and mediated searching. Part 4: Cognitive styles in information seeking," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 728–735, 2002.

[102] J. Foster, "Collaborative information seeking and retrieval," *Annual Review of Information Science and Technology*, vol. 40, pp. 329–356, 2006.

[103] L. Freund and E. Toms, "Revisiting informativeness as a process measure for information interaction," in *Proceedings of the Web Information-Seeking and Interaction* (*WISI*) *Workshop at the 30th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval (*SIGIR '07*), pp. 33–36, Amsterdam, The Netherlands, 2007.

[104] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human–system communication: An analysis and a solution," *Communications of the ACM*, vol. 30, pp. 964–971, 1987.

[105] R. M. Furr and V. R. Bacharach, *Psychometrics: An Introduction*. Sage Publications, Inc, 2007.

[106] J. A. Ghani and S. P. Deshpande, "Task characteristics and the experience of optimal flow in Human–Computer interaction," *Journal of Psychology*, vol. 128, pp. 381–391, 1994.

[107] J. A. Ghani, R. Supnick, and P. Rooney, "The experience of flow in computer-mediated and in face-to-face groups," in *Proceedings of International Conference on Information Systems* (*ICIS 1991*), pp. 229–237, New York, NY, 1991.

[108] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine, 1967.

[109] A. Göker and H. Myrhaug, "Evaluation of a mobile information system in context," *Information Processing and Management*, vol. 44, pp. 39–65, 2008.

[110] F. J. Gravetter and L. B. Wallnau, *Statistics for the Behavioral Sciences*. Thomson Learning, Fifth ed., 1999.

[111] C. Grimes, D. Tang, and D. M. Russell, "Query logs alone are not enough," in *Proceedings of the Workshop on Query Log Analysis: Social and Technology Challenges at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[112] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The validity of the stimulated retrospective think-aloud method as measured by eye tracking," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1253–1262, Montreal, Canada, 2006.

[113] J. P. Guilford, *Fundamental Statistics in Psychology and Education*. New York: McGraw Hill, 1956.

[114] J. Gwizdka, "Revisiting search task difficulty: Behavioral and individual difference measures," *Proceedings of the 71th Annual Meeting of the American Society for Information Science and Technology* (*ASIS and T*), 2008.

[115] D. F. Haas and D. H. Kraft, "Experimental and quasi-experimental designs for research in information science," *Information Processing and Management*, vol. 20, pp. 229–237, 1984.

[116] K. Halttunen and K. Järvelin, "Assessing learning outcomes in two information retrieval learning environments," *Information Processing and Management*, vol. 41, pp. 949–972, 2005.

[117] P. A. Hancock and N. Meshkati, *Human Mental Workload*. The Netherlands: Elsevier Science Publishers, 1988.

[118] S. Hansell, AOL removes search data on vase group of Web users, New York Times, Friday, March 14, 2008. Business Section. http://query.nytimes.com/gst/fullpage.html?res=9504E5D81E3FF93BA3575BC0A9609C8B63, 2006.

[119] S. Harada, M. Naaman, Y. J. Song, Q.-Y. Wang, and A. Paepcke, "Lost in memories: Interacting with photo collections on PDAs," in *Proceedings of*

the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (*JCDL '04*), pp. 325–333, Tuscon, AZ, 2004.

[120] D. K. Harman, "Introduction to special issue on evaluation issues in information retrieval," *Information Processing and Management*, vol. 28, pp. 439–440, 1992.

[121] D. K. Harman, "The TREC test collection," in *TREC: Experiment and Evaluation in Information Retrieval*, (E. M. Voorhees and D. K. Harman, eds.), pp. 21–52, Cambridge, MA: MIT Press, 2005.

[122] S. G. Hart and L. E. Staveland, "Development of a NASA-TLX (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, (P. Hancock and N. Meshkati, eds.), pp. 139–183, The Netherlands: Elsevier Science Publishers, 1988.

[123] S. P. Harter and C. A. Hert, "Evaluation of information retrieval systems: Approaches, issues and methods," *Annual Review of Information Science and Technology*, vol. 32, pp. 3–94, 1997.

[124] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and ergonomic quality aspects determine a software's appeal," *Proceedings of the ACM Conference on Human Factors in Computing Systems* (*CHI '00*), pp. 201–208, 2000.

[125] D. Hawking, P. Bailey, and N. Craswell, "Efficient and flexible search using text and metadata," *CSIRO Mathematical and Information Sciences Tech Report, 2000-83*, available online at http://es.csiro.au/pubs/hawking_tr00b.pdf, 2000.

[126] M. D. Heine, "Simulation, and simulation experiments," in *Information Retrieval Experiment*, (K. Spärck-Jones, ed.), pp. 179–198, London, UK: Butterworths and Co. Ltd, 1981.

[127] W. Hersh, "Evaluating interactive question answering," in *Advances in Open Domain Question Answering*, (T. Strzalkowski and S. Harabagiu, eds.), pp. 431–455, Dordrecht, The Netherlands: Springer, 2006.

[128] W. Hersh and P. Over, "Introduction to a special issue on interactivity at the Text Retrieval Conference (TREC)," *Information Processing and Management*, vol. 37, pp. 365–367, 2001.

[129] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, "Do batch and user evaluations give the same results?," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '00*), pp. 17–24, Athens, Greece, 2000.

[130] W. R. Hersh, D. L. Elliot, D. H. Hickam, S. L. Wolf, A. Molnar, and C. Leichtenstein, "Towards new measures of information retrieval evaluation," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '95*), pp. 164–170, Seattle, WA, 1995.

[131] S. Hirsh and J. Dinkelacker, "Seeking information in order to produce information: An empirical study at Hewlett Packard Labs," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 807–817, 2004.

[132] K. Hornbaek, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Human–Computer Studies*, vol. 64, pp. 79–102, 2005.

[133] K. Hornbæk and E. L.-C. Law, "Meta-analysis of correlations among usability measures," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 617–626, San Jose, CA, 2007.

[134] I. Hsieh-Yee, "Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers," *Journal of the American Society for Information Science*, vol. 44, pp. 161–174, 1993.

[135] M. Huang and H. Wang, "The influence of document presentation order and number of documents judged on users' judgments of relevance," *Journal of the American Society for Information Science*, vol. 55, pp. 970–979, 2004.

[136] G. Iachello and J. Hong, "End-user privacy in human–computer interaction," *Foundations and Trends in Human–Computer Interaction*, vol. 1, pp. 1–137, 2007.

[137] P. Ingwersen, *Information Retrieval Interaction*. London: Taylor Graham, 1992.

[138] P. Ingwersen, "Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory," *Journal of Documentation*, vol. 52, pp. 3–50, 1996.

[139] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht, The Netherlands: Springer, 2005.

[140] P. Ingwersen and P. Willett, "An introduction to algorithmic and cognitive approaches for information retrieval," *Libri*, vol. 45, pp. 160–177, 1995.

[141] International Standards Office, *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part II, Guidance on Usability* (ISO 9241-11:1998). 1998.

[142] E. Jacob, "Qualitative research traditions: A review," *Review of Educational Research*, vol. 57, pp. 1–50, 1987.

[143] R. J. K. Jacob and K. S. Karn, "Eye tracking in human–Computer interaction and usability research: Ready to deliver the promises (section commentary)," in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, (J. Hyona, R. Radach, and H. Deubel, eds.), pp. 573–605, Amsterdam, Elsevier Science, 2003.

[144] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it," *Library and Information Science Research*, vol. 28, pp. 407–432, 2006.

[145] B. J. Jansen, R. Ramadoss, M. Zhang, and N. Zang, "Wrapper: An application for evaluating exploratory searching outside of the lab," in *Workshop on Evaluating Exploratory Search Systems at the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '06)*, Seattle, WA, 2006.

[146] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, pp. 248–263, 2005.

[147] K. Järvelin, "An analysis of two approaches in information retrieval: From frameworks to study designs," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 971–986, 2007.

[148] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd ACM SIGIR Conference on Research and Development of Information Retrieval* (*SIGIR '00*), pp. 41–48, Athens, Greece, 2000.

[149] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems* (*TOIS*), vol. 20, pp. 422–446, 2002.

[150] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen, "Discounted cumulated gain based evaluation of multiple-query IR sessions," in *Proceedings of the 30th European Conference on Information Retrieval*, Glasgow, Scotland, 2008.

[151] H. R. Jex, "Measuring mental workload: Problems, process and promises," in *Human Mental Workload*, (P. Hancock and N. Meshkati, eds.), pp. 5–39, The Netherlands: Elsevier Science Publishers, 1988.

[152] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '05*), pp. 154–161, Salvador, Brazil, 2005.

[153] H. Joho, D. Hannah, and J. M. Jose, "Comparing collaborative and independent search in a recall-oriented task," in *Proceedings of the 2nd IIiX Symposium on Information Interaction in Context*, pp. 89–96, London, UK, 2008.

[154] H. Joho and J. M. Jose, "Effectiveness of additional representations for the search result presentation on the Web," *Information Processing and Management*, vol. 44, pp. 226–241, 2008.

[155] W. Jones and J. Teevan, *Personal Information Management*. Seattle, WA: University of Washington Press, 2007.

[156] C.-A. Julien, J. E. Leide, and F. Bouthillier, "Controlled user evaluations of information visualization interfaces for text retrieval: Literature review and meta-analysis," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 1012–1024, 2008.

[157] H. Julien, L. E. F. McKechnie, and S. Hart, "Affective issues in library and information science systems work: A content analysis," *Library and Information Science Research*, vol. 27, pp. 453–466, 2005.

[158] M. Käki and A. Aula, "Controlling the complexity in comparing search user interfaces via user studies," *Information Processing and Management*, vol. 44, pp. 82–91, 2008.

[159] J. Kalgren and K. Franzen, "Verbosity and interface design," Retrieved on 01 February 2008 at http://www.ling.su.se/staff/franzen/irinterface.html, 1997.

[160] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, "Proceedings of the 2008 ACM workshop on research advances in large digital book repositories," in *Proceedings of the Conference on Information and Knowledge Management* (*CIKM '08*), Napa, CA, 2008.

[161] P. B. Kantor, "Evaluation of and feedback in information storage and retrieval systems," *Annual Review of Information Science and Technology*, vol. 17, pp. 99–120, 1982.

[162] J. Kekäläinen and K. Järvelin, "Using graded relevance assessments in IR evaluation," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 1120–1129, 2002.

[163] M. Kellar, C. Watters, and M. Shepherd, "A field study characterizing Web-based information-seeking tasks," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 999–1018, 2007.

[164] D. Kelly and X. Fu, "Elicitation of term relevance feedback: An investigation of term source and context," in *Proceedings of the 29th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 453–460, Seattle, WA, 2006.

[165] D. Kelly, J. Harper, and B. Landau, "Questionnaire mode effects in interactive information retrieval experiments," *Information Processing and Management*, 2008.

[166] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, "User-centered evaluation of interactive question answering systems," in *Proceedings of the Workshop on Interactive Question Answering at the Human Language Technology Conference* (*HLT-NAACL '06*), pp. New York, NY, 2006.

[167] D. Kelly and J. Lin, "Overview of the TREC 2006 ciQA task," *SIGIR Forum*, vol. 41, pp. 107–116, 2007.

[168] D. Kelly, C. Shah, C. R. Sugimoto, E. W. Bailey, R. A. Clemens, A. K. Irvine, N. A. Johnson, W. Ke, S. Oh, A. Poljakova, M. A. Rodriguez, M. G. van Noord, and Y. Zhang, "Effects of performance feedback on users' evaluations of an interactive IR system," in *Proceedings of the 2nd Symposium on Information Interaction in Context* (*IIiX*), pp. 75–82, London, UK, 2008.

[169] D. Kelly, N. Wacholder, R. Rittman, Y. Sun, P. Kantor, S. Small, and T. Strzalkowski, "Using interview data to identify evaluation criteria for interactive, analytical question answering systems," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1032–1043, 2007.

[170] J. Kim, "Task as a predictable indicator for information seeking behavior on the Web," Ph.D. Dissertation, Rutgers University, 2006.

[171] K.-S. Kim and B. Allen, "Cognitive and task influences on Web searching behavior," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 109–119, 2001.

[172] S. Kim and D. Soergel, "Selecting and measuring task characteristics as independent variables," *Proceedings of the American Society for Information Science and Technology Conference*, vol. 42, 2006.

[173] D. W. King, "Design and evaluation of information systems," *Annual Review of Information Science and Technology*, vol. 3, pp. 61–103, 1968.

[174] C. C. Kuhlthau, *Seeking Meaning: A Process Approach to Library and Information Services*. Norwood, NJ: Ablex, 1993.

[175] E. Lagergren and P. Over, "Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment," in *Proceedings of the 21st Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, pp. 164–172, Melbourne, Australia, 1998.

[176] B. Larsen, S. Malik, and A. Tombros, "The interactive track at INEX2005," in *INEX 2005*, (N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds.), pp. 398–410, Berlin: Springer, 2006.

[177] Y. Li and N. J. Belkin, "A faceted approach to conceptualizing tasks information seeking," *Information Processing and Management*, vol. 44, pp. 1822–1837, 2008.

[178] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[179] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger, "What makes a good answer? The role of context in question answering," in *Proceedings of the 9th IFIP TC13 International Conference on Human–Computer Interaction* (*INTERACT 2003*), (M. Rauterberg, M. Menozzi, and J. Wesson, eds.), Zurich, Switzerland, 2003.

[180] J. Lin and M. Smucker, "How do users find things with PubMed? Towards automatic utility evaluation with user simulations," in *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 08*), pp. 19–26, Singapore, Malaysia, 2008.

[181] S. J. Lin and N. J. Belkin, "Modeling multiple information seeking episodes," in *Proceedings of the Annual Conference of the American Society for Information Science* (*ASIS '00*), pp. 133–146, USA, 2000.

[182] S. Littlejohn, *Theories of Human Communication*. Belmont, CA: Wadsworth, 1992.

[183] I. Lopatovska and H. B. Mokros, "Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts," *Information Processing and Management*, vol. 44, pp. 92–104, 2008.

[184] L. Lorigo, M. Haridasan, H. Brynjarsdottir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan, "Eye tracking and online search: Lessons learned and challenges ahead," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 1041–1052, 2008.

[185] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay, "The influence of task and gender on search and evaluation behavior using Google," *Information Processing and Management*, vol. 42, pp. 1123–1131, 2006.

[186] R. M. Losee, "Evaluating retrieval performance given database and query characteristics: Analytical determination of performance surfaces," *Journal of the American Society for Information Science*, vol. 47, pp. 95–105, 1996.

[187] H. P. Luhn, "A business intelligence system," in *H. P. Luhn: Pioneer of Information Science. Selected Works*, (C. K. Shultz, ed.), pp. 132–139, NY: Spartan Books, 1958.

[188] A. M. Lund, "Measuring usability with the USE Questionnaire," *Usability and User Experience*, vol. 8, Available online: http://www.stcsig.org/ usability/newsletter/0110_measuring_with_use.html, 2001.

[189] W. E. Mackay, "Ethics, lies and videotape," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 138–145, Denver, CO, 1995.

[190] G. Marchionini, "Exploratory search: From finding to understanding," *Communications of the ACM*, vol. 49, pp. 41–46, 2006a.

[191] G. Marchionini, "Toward human–Computer information retrieval," *Bulletin of the American Society for Information Science*, online: http://www.asis.org/Bulletin/Jun-06/marchionini.html (retrieved November 7, 2008), 2006b.

[192] G. Marchionini and G. Crane, "Evaluating hypermedia and learning: Methods and results from the Perseus Project," *ACM Transactions on Information Systems*, vol. 12, pp. 5–34, 1994.

[193] J. A. Maxwell, *Qualitative Research Design: An Interactive Approach*. CA: Sage Publications, 1996.

[194] D. A. Michel, "What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information," *Journal of the American Society for Information Science*, vol. 45, pp. 498–514, 1994.

[195] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: A Sourcebook of New Methods*. Newbury Park: Sage, 1984.

[196] L. I. Millett, B. Friedman, and E. Felten, "Cookies and web browser design: Toward realizing informed consent online," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 46–52, Seattle, WA, 2001.

[197] S. Mizzaro, "Relevance: The whole history," *Journal of the American Society for Information Science*, vol. 48, pp. 810–832, 1997.

[198] J. Morahan-Martin, "Males, females, and the Internet," in *Psychology and the Internet: Intrapersonal, Interpersonal, and Transpersonal Implications*, (J. Gackenback, ed.), pp. 169–197, San Diego: Academic Press, 1998.

[199] M. R. Morris and E. Horvitz, "SearchTogether: An interface for collaborative Web search," in *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*, pp. 3–12, New York, NY, 2007.

[200] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*. Mahway, NJ: Lawrence Erlbaum Associates Inc., Publishers, Second ed., 2003.

[201] K. A. Neuendorf, *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications, 2002.

[202] J. Nielsen, "Usability 101: Introduction to usability," *Jakob Nielsen's Alertbox*, Retrieved May 05, 2008 from http://www.useit.com/alertbox/20030825.html, 2003.

[203] J. Nielsen and J. Levy, "Measuring usability: Preference vs performance," *Communications of the ACM*, vol. 37, pp. 66–75, 1994.

[204] D. Oard, "Evaluating interactive cross-language information retrieval: Document selection," in *Proceedings of the 1st Cross-Language Evaluation Forum*, Lisbon, Portugal, 2000.

[205] H. O'Brien and E. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 938–955, 2008.

[206] H. L. O'Brien, E. G. Toms, E. K. Kelloway, and E. Kelley, "Developing and evaluating a reliable measure of user engagement," in *Proceedings of the American Society for Information Science and Technology*, Columbus, Ohio, 2008.

[207] S. L. Payne, *The Art of Asking Questions*. Princeton, NJ: Princeton University Press, 1951.

[208] E. J. Pedhazur and L. P. Schmelkin, *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[209] T. A. Peters, "The history and development of transaction log analysis," *Library Hi Tech*, vol. 11, pp. 41–66, 1993.

[210] T. A. Peters, M. Kurth, P. Flaherty, B. Sandore, and N. K. Kaske, "An introduction to the special section on transaction log analysis," *Library Hi Tech*, vol. 11, pp. 38–40, 1993.

[211] D. Petrelli, "On the role of user-centred evaluation in the advancement of interactive information retrieval," *Information Processing and Management*, vol. 44, pp. 22–38, 2008.

[212] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[213] E. M. Pilke, "Flow experiences in information technology use," *International Journal of Human–Computer Studies*, vol. 61, pp. 347–357, 2004.

[214] D. J. Pittenger, "Internet research: An opportunity to revisit classic ethical problems in behavioral research," *Ethics and Behavior*, vol. 13, pp. 45–60, 2003.

[215] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, vol. 88, pp. 879–903, 2003.

[216] R. J. Riding and I. Cheema, "Cognitive styles — An overview and integration," *Education Psychology*, vol. 11, pp. 193–215, 1991.

[217] S. Y. Rieh, "Judgment of information quality and cognitive authority in the Web," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 145–161, 2002.

[218] S. E. Robertson, "The probability ranking principle in IR," *Journal of Documentation*, vol. 33, pp. 294–304, 1977.

[219] S. E. Robertson, "On GMAP — And other transformations," in *Proceedings of the 15th ACM international Conference on information and Knowledge Management* (*CIKM'06*), pp. 78–83, Arlington, VA, 2006.

[220] S. E. Robertson, "On the history of evaluation in IR," *Journal of Information Science*, vol. 34, pp. 439–456, 2008.

[221] S. E. Robertson and M. M. Hancock-Beaulieu, "On the evaluation of IR systems," *Information Processing and Management*, vol. 28, pp. 457–466, 1992.

[222] K. Rodden and X. Fu, "Exploring how mouse movements relate to eye movements on Web search results pages," in *Proceedings of the Web Information-Seeking and Interaction* (*WISI*) *Workshop at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '07*), pp. 29–32, Amsterdam, The Netherlands, 2007.

[223] M. E. Rorvig, "Psychometric measurement and information retrieval," *Annual Review of Information Science and Technology*, vol. 23, pp. 157–189, 1988.

[224] I. Ruthven, "Integrating approaches to relevance," in *New Directions in Cognitive Information Retrieval*, (A. Spink and C. Cole, eds.), pp. 61–80, Netherlands: Springer, 2005.

[225] I. Ruthven, "Interactive information retrieval," *Annual Review of Information Science and Technology*, vol. 42, pp. 43–91, 2008.

[226] I. Ruthven, M. Baillie, and D. Elsweiler, "The relative effects of knowledge, interest and confidence in assessing relevance," *Journal of Documentation*, vol. 63, pp. 482–504, 2007.

[227] I. Ruthven, P. Borlund, P. Ingwersen, N. J. Belkin, A. Tombros, and P. Vakkari in *Proceedings of the 1st International Conference on Information Interaction in Context*, Copenhagen, Denmark, 2006.

[228] K. J. Ryan, J. V. Brady, R. E. Cooke, D. I. Height, A. R. Jonsen, P. King, K. Lebacqz, D. W. Louisell, D. Seldin, E. Stellar, and R. H. Turtle, "The Belmont Report," Available at http://www.hhs.gov/ohrp/humansubjects/ guidance/belmont.htm, 1979.

[229] G. Salton, "Evaluation problems in interactive information retrieval," *Information Storage and Retrieval*, vol. 6, pp. 29–44, 1970.

[230] G. Salton, "The state of retrieval system evaluation," *Information Processing and Management*, vol. 28, pp. 441–449, 1992.

[231] T. Saracevic, "Quo vadis test and evaluation," in *Proceedings of the Annual Meeting of the American Documentation Institute, 4*, pp. 100–104, New York, NY, 1967.

[232] T. Saracevic, "Relevance: A review of and a framework for the thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, pp. 321–343, 1975.

[233] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development of Information Retrieval*, pp. 138–146, Seattle, WA, 1995.

[234] T. Saracevic, "The stratified model of information retrieval interaction: Extension and applications," *Proceedings of the American Society for Information Science Conference*, vol. 34, pp. 313–327, 1997.

[235] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1915–1933, 2007a.

[236] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 2126–2144, 2007b.

[237] T. Saracevic and P. B. Kantor, "A study of information seeking and retrieving. II. Users, questions and effectiveness," *Journal of the American Society for Information Science*, vol. 39, pp. 177–196, 1988a.

[238] T. Saracevic and P. B. Kantor, "A study of information seeking and retrieving. III. Searchers, searches and overlap," *Journal of the American Society for Information Science*, vol. 39, pp. 197–216, 1988b.

[239] T. Saracevic, P. B. Kantor, A. Y. Chamis, and D. Trivison, "A study of information seeking and retrieving: Part 1, background and methodology,"

Journal of the American Society for Information Science, vol. 39, pp. 161–176, 1988.

[240] P. A. Savage-Knepshield and N. J. Belkin, "Interaction in information retrieval: Trends over time," *Journal of the American Society for Information Science*, vol. 50, pp. 1067–1082, 1999.

[241] R. Savolainen, "Everyday life information seeking: Approaching information seeking in the context of way of life," *Library and Information Science Research*, vol. 17, pp. 259–294, 1995.

[242] R. Siatri, "The evolution of user studies," *Libri*, vol. 49, pp. 132–141, 1999.

[243] C. L. Smith and P. B. Kantor, "User adaptation: Good results from poor systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 147–154, Singapore, 2008.

[244] P. Solomon, "Discovering information in context," *Annual Review of Information Science and Technology*, vol. 36, pp. 229–264, 2002.

[245] E. Sormunen, "Liberal relevance criteria of TREC: Counting on negligible documents?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '02*), pp. 324–330, Tampere, Finland, 2002.

[246] K. Spärck-Jones, *Information Retrieval Experiment*. London, UK: Butterworths and Co. Ltd, 1981.

[247] A. Spink, "Multiple search session model of end-user behavior: An exploratory study," *Journal of the American Society for Information Science*, vol. 47, pp. 603–609, 1996.

[248] A. Spink, "Study of interactive feedback during mediated information retrieval," *Journal of the American Society for Information Science*, vol. 48, pp. 382–394, 1997.

[249] A. Spink and H. Greisdorf, "Regions and levels: Measuring and mapping users' relevance judgments," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 161–173, 2001.

[250] A. Spink and R. M. Losee, "Feedback in information retrieval," *Annual Review of Information Science and Technology*, vol. 31, pp. 33–78, 1996.

[251] A. Spink, H. C. Ozmutlu, and S. Ozmutlu, "Multitasking information seeking and searching processes," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 639–652, 2002.

[252] S. R. Stern, "Encountering distressing information in online research: A consideration of legal and ethical responsibilities," *New Media and Society*, vol. 5, pp. 249–266, 2003.

[253] A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. CA: Sage Publications, 1990.

[254] L. T. Su, "Evaluation measures for interactive information retrieval," *Information Processing and Management*, vol. 28, pp. 503–516, 1992.

[255] W. Sugar, "User-centered perspectives of information retrieval research and analysis methods," *Annual Review of Information Science and Technology*, vol. 30, pp. 77–109, 1995.

[256] SUMI Questionnaire. (retrieved November 05, 2008), http://www.ucc.ie/hfrg/questionnaires/sumi/index.html.

[257] Y. Sun and P. Kantor, "Cross-evaluation: A new model for information system evaluation," *Journal of American Society for Information Science and Technology*, vol. 56, pp. 614–628, 2006.

[258] J. Tague, "Informativeness as an ordinal utility function for information retrieval," *SIGIR Forum*, vol. 21, pp. 10–17, 1987.

[259] J. Tague and R. Schultz, "Evaluation of the user interface in an information retrieval system: A model," *Information Processing and Management*, vol. 25, pp. 377–389, 1988.

[260] J. M. Tague, "The pragmatics of information retrieval experimentation," in *Information Retrieval Experiment*, (K. S. Jones, ed.), pp. 59–104, London, UK: Butterworths and Co. Ltd, 1981.

[261] J. M. Tague-Sutcliffe, "The pragmatics of information retrieval experimentation, revisted," *Information Processing and Management*, vol. 28, pp. 467–490, 1992a.

[262] J. M. Tague-Sutcliffe, "Measuring the informativeness of a retrieval process," in *Proceedings of the 15th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '92*), pp. 23–36, Copenhagen, Denmark, 1992b.

[263] J. M. Tague-Sutcliffe, *Measuring Information: An Information Services Perspective*. San Diego, California: Academic Press, 1995.

[264] J. M. Tague-Sutcliffe, "Some perspectives on the evaluation of information retrieval systems," *Journal of the American Society for Information Science*, vol. 47, pp. 1–3, 1996.

[265] R. Tang, M. Shaw, and J. L. Vevea, "Towards the identification of the optimal number of relevance categories," *Journal of the American Society for Information Science*, vol. 50, pp. 254–264, 1999.

[266] A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio, "Relationships between categories of relevance criteria and stage in task completion," *Information Processing and Management*, vol. 43, pp. 1071–1084, 2007.

[267] R. S. Taylor, "Question negotiation and information seeking in libraries," *College and Research Libraries*, vol. 29, pp. 178–194, 1968.

[268] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, "The perfect search engine is not enough: A study of orienteering behavior in directed search," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (*SIGCHI '04*), pp. 415–422, Vienna, Austria, 2004.

[269] J. Teevan, S. T. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '05*), pp. 449–456, Salvador, Brazil, 2005.

[270] P. Thomas and D. Hawking, "Evaluation by comparing result sets in context," in *Proceedings of the 15th Annual Conference on Information and Knowledge Management* (*CIKM '06*), pp. 94–101, Arlington, VA, 2006.

[271] A. Tombros, I. Ruthven, and J. M. Jose, "How users assess Web pages for information seeking," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 327–344, 2005.

[272] E. G. Toms, L. Freund, and C. Li, "W*i*IRE: The Web interactive information retrieval experimentation system prototype," *Information Processing and Management*, vol. 40, pp. 655–675, 2004.

[273] E. G. Toms, H. L. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. MacNutt, "Task effects on interactive search: The query factor," *Proceedings of INEX 2007*, pp. 359–372, 2007.

[274] R. Tourangeau, L. J. Rips, and K. Rasinski, *The Psychology of Survey Response*. New York, NY: Cambridge University Press, 2000.

[275] A. Turpin and W. Hersh, "Why batch and user evaluations do not give the same results," in *Proceedings of the 24th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '01*), pp. 225–231, New Orleans, LA, 2001.

[276] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 11–18, Seattle, WA, 2006.

[277] H. Turtle, W. D. Penniman, and T. Hickey, "Data entry/display devices for interactive information retrieval," *Annual Review of Information Science and Technology*, vol. 16, pp. 55–83, 1981.

[278] D. J. Urquhart, "The distribution and use of scientific and technical information," *The Royal Society Scientific Information Conference*, pp. 408–419, 1948.

[279] P. Vakkari, "Task-based information searching," *Annual Review of Information Science and Technology*, vol. 37, pp. 413–464, 2003.

[280] P. Vakkari, "Changes in search tactics and relevance judgments when preparing a research proposal: A summary of the findings of a longitudinal study," *Information Retrieval*, vol. 4, pp. 295–310, 2004.

[281] P. Vakkari and N. Hakala, "Changes in relevance criteria and problem stages in task performance," *Journal of Documentation*, vol. 56, pp. 540–562, 2000.

[282] P. Vakkari and K. Järvelin, "Explanation in information seeking and retrieval," in *New Directions in Cognitive Information Retrieval*, (A. Spink and C. Cole, eds.), pp. 113–138, Berlin: Springer, The Information Retrieval Series, 2005.

[283] P. Vakkari and E. Sormunen, "The influence of relevance levels on the effectiveness of interactive information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 963–969, 2004.

[284] A. Veerasamy and N. J. Belkin, "Evaluation of a tool for visualization of information retrieval results," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 85–92, Zurich, Switzerland, 1996.

[285] A. Veerasamy and R. Heikes, "Effectiveness of a graphical display of retrieval results," *SIGIR Forum*, vol. 31, pp. 236–245, 1997.

[286] M. Viswanathan, *Measurement Error and Research Design*. Sage Publications, Inc, 2005.

[287] E. M. Voorhees, "On test collections for adaptive information retrieval," *Information Processing and Management*, vol. 44, pp. 1879–1885, 2008.

[288] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 2005.

[289] J. B. Walther, "Research ethics in Internet-enabled research: Human subjects issues and methodological myopia," *Ethics and Information Technology*, vol. 4, pp. 205–216, 2002.

[290] P. Wang, "Methodologies and methods for user behavior research," *Annual Review of Information Science and Technology*, vol. 34, pp. 53–99, 1999.

[291] L. Wen, I. Ruthven, and P. Borlund, "The effects on topic familiarity on online search behaviour and use of relevance criteria," in *Proceedings of the 28th European Conference in Information Retrieval* (*ECIR 2006*), London, UK, 2006.

[292] R. W. White, M. Bilenko, and S. Cucerzan, "Studying the use of popular destinations to enhance web search interaction," in *Proceedings of the 30th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '07*), pp. 159–166, Amsterdam, The Netherlands, 2007.

[293] R. W. White, I. Ruthven, J. M. Jose, and C. J. van Rijsbergen, "Evaluating implicit feedback models using searcher simulations," *ACM Transactions on Information Systems*, vol. 23, pp. 325–261, 2005.

[294] B. M. Wildemuth, *Applications of Social Research Methods to Questions in Information and Library Science*. Libraries Unlimited, (in press).

[295] B. M. Wildemuth, M. Yang, A. Hughes, R. Gruss, G. Geisler, and G. Marchionini, "Access via features versus access via transcripts: User performance and satisfaction," *TREC VID 2003 Notebook Paper*, 2003.

[296] F. Williams, *Reasoning with Statistics: How to Read Quantitative Research*. Orlando, Florida: Holt, Rinehart and Winston, Inc, Fourth ed., 1992.

[297] T. D. Wilson, "On user studies and information needs," *Journal of Documentation*, vol. 37, pp. 3–15, 1981.

[298] L. Xiong and E. Agichtein, "Towards privacy-preserving query log publishing," in *Proceedings of the Workshop on Query Log Analysis: Social and Technology Challenges at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[299] W. Yuan and C. T. Meadow, "A study of the use of variables in information retrieval user studies," *Journal of the American Society for Information Science*, vol. 50, pp. 140–150, 1999.

[300] P. Zhang, L. Plettenberg, J. L. Klavans, D. W. Oard, and D. Soergel, "Task-based interaction with an integrated multilingual, multimedia information system: A formative evaluation," in *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries* (*JCDL '07*), pp. 117–126, Vancouver, BC, 2007.

[301] X. Zhang, "Collaborative relevance judgments: A group consensus method for evaluating user search performance," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 220–231, 2002.