Experimental Methods in IIR: The Tension between Rigour and Ethics in Studies Involving Users with Dyslexia

Gerd Berget
Dep. of Archivistics, Library and Information Science
Oslo Metropolitian University
Oslo, Norway
gerd.berget@oslomet.no

Andrew MacFarlane
Centre for HCI Design
City, University of London
London, UK
a.macfarlane@city.ac.uk

ABSTRACT

Designing user studies in the interactive information retrieval (IIR) paradigm on people with impairments may sometimes require different methodological considerations than for other users. Consequently, there may be a tension between what the community regards as being a rigorous methodology against what researchers can do ethically with their users. This paper discusses issues to consider when designing IIR studies involving people with dyslexia, such as sampling, informed consent and data collection. The conclusion is that conducting user studies on participants with dyslexia requires special considerations at all stages of the experimental design. The purpose of this paper is to raise awareness and understanding in the research community about experimental methods involving users with dyslexia, and addresses researchers, as well as editors and reviewers. Several of the issues raised do not only apply to people with dyslexia, but have implications when researching other groups, for instance elderly people and users with learning, cognitive, sensory or motor impairments.

CCS CONCEPTS

Information systems→Information retrieval→Users and interactive retrieval→Search interfaces • Human-centred computing→Accessibility→Empirical studies in accessibility

KEYWORDS

Interactive information retrieval; User studies; Dyslexia; Impairments

ACM Reference format:
Gerd Berget, and Andrew MacFarlane. 2018. Experimental Methods in IIR: The Tension Between Rigour and Ethics in Studies Involving Users with Dyslexia. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR'19), March 10-14, 2018, Glasgow, UK.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3295750.3298939

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298939

1 INTRODUCTION

Over the past two decades, an increasing number of user studies on information retrieval have addressed people with impairments [21]. In the last ten years or so, we have conducted several studies within the field of IIR with a focus on users with dyslexia [3-5; 28-30]. Through these studies we have experienced many methodological issues where there is a mismatch between rigorous methodology and what you can actually do in experiments regarding people with impairments. In this perspective paper we will discuss some of these issues, and suggest alternative methodological approaches and considerations. The main focus is on laboratory evaluations of the interaction between users and systems. However, several issues apply to other research methods, for instance sampling, informed consent, how to document diagnoses and reporting results. Our aim is to outline the best practice for undertaking IIR experiments involving users with dyslexia, in the light of particular ethical issues which arise with this user group.

Dyslexia is a cognitive impairment with an estimated prevalence of around 7% of the world population [42]. Although dyslexia is mainly characterised by reading and writing difficulties, the cognitive profile is often quite complex and typically includes impaired short-term memory [41], reduced rapid naming skills [39] and concentration difficulties, especially when focusing on a task for a long time [33]. Moreover, it has been reported that people with dyslexia have challenges with mental organisation [24].

Reduced self-esteem is another well-documented characteristic related to dyslexia, in addition to anxiety and fear of failure, among others caused by negative experiences at school [40]. Comorbidity with other diagnoses is also quite common. A total of 18-20% of the dyslexic population also have ADHD or ADD [15]. Prevalence with other diagnoses is also frequently reported, such as the writing impairment dysgraphia [37], the mathematical disorder dyscalculia [27] and dyspraxia, an impairment in the organization of movement [16].

Detailed knowledge about the cognitive and physical abilities of the participants is a very important aspect of designing studies involving users with impairments. For instance, the cognitive profile of people with dyslexia affects the ability to read and understand written information and the capability to complete comprehensive tasks in long-lasting sessions. We will revisit these challenges and present alternative procedures for

conducting experiments involving this user group in the following sections.

The paper is structured as follows: firstly, we set the scene by outlining the issue of vulnerable groups and ethical concerns. In the light of these we address standard experimental methods in IIR, with a focus on the topics that are most relevant for users with impairments, namely sampling, informed consent, method of data collection and reporting results, outlining problems at each stage. We then suggest revised experimental methods in the areas addressed. A set of recommendations for researchers undertaking work involving users with dyslexia is then provided, together with a concluding statement on prospects in the field as a whole.

2 VULNERABLE USER GROUPS

Many studies today on users with impairments are conducted within the universal design paradigm, where the main goal is to design systems that can be used by all people, regardless of cultural background, gender, age and physical and mental abilities [47]. However, while there is an increasing awareness that people with impairments need to be included in research, there has been little focus on the experimental design of such studies.

Much IIR experimentation has focused on the general population, with very little work until recently on users with cognitive issues such as dyslexia [3-5; 28-30]. Therefore, little attention has been paid to the differences these users have with the general population and the impact this impairment may have on the result of any experiment conducted. In order to address the issues of concern regarding users with specific impairments, more focus on ethical issues is required. This is because such users (including people with dyslexia) are regarded as being vulnerable. For people with dyslexia, vulnerability takes a number of forms including a lack of self-esteem [35; 49], a tendency to become tired after significant intellectual effort [33] and anxiety [9; 40].

Ethical panels require researchers to give extra consideration to vulnerable groups, over and above what would be expected with the general population. For example, at City, University of London researchers are specifically asked to identify the given vulnerable group and identify a specific strategy to ensure their well-being, they do not feel pressured to take part and are free to withdraw at any time from the study [10]. Approval for the study is only granted if the ethics panel consider these issues fully addressed and the participants are treated in an appropriate ethical manner. For users with dyslexia IIR researchers have a clear responsibility to ensure that users' self-esteem is not reduced further, they do not put together an experiment which will tire out the users or make them feel anxious whilst undertaking it. In this light we examine the standard methods for experimentation in IIR.

3 EXPERIMENTAL METHODS IN IIR

In IIR studies, researchers address the interaction between users and information systems. A challenge within the field of IIR has been to create experimental situations that consider individual factors, while at the same time makes it possible to establish causal relationships. According to Kelly [25], there has been little guidance on experimental frameworks and there has been a lack of well-established methods in IIR.

The experimental setting often used for the evaluation of IIR systems typically consists of three components: users, information needs and relevance judgements [6]. An important topic for IIR is whether people can actually use the system in question. We argue that studies on people with impairments need to follow the cognitive user-centred approach as described by Borlund [6], and not a system-driven approach. Users with impairments have specific issues that can only be addressed through a user centred approach, and the focus of experiments needs to be on the users rather than the system.

3.1 Sampling

Sampling is much discussed in methodological literature. According to standard textbooks, the sampling frame, a complete list of sampling units, is an important starting point when recruiting participants to ensure a representative population [14]. However, lists of people with impairments are not accessible to researchers, for ethical reasons, since such diagnoses represent highly sensitive information. Ethics panels will not allow researchers to use prior diagnosis without explicit consent from the user (which may not be forthcoming as the user may not wish to disclose sensitive personal information). Consequently, the sampling process must rely on volunteering participants, making it challenging to assemble a balanced sample.

Classic experimental design often involves two comparable groups: one experimental group and one control group, where only the independent variable should differ between the two groups. In the context of dyslexia, the control group should therefore consist of people without this diagnosis. According to Mortimore and Crozier [33] adults may not always be aware of their dyslexia. Consequently, researchers may potentially include people with undiagnosed dyslexia in the control group, weakening the validity of the experimental design.

3.2 Informed consent

According to Frankfort-Nachmias, Nachmias and DeWaard [14], informed consent is the most important tool to ensure the welfare and rights of participants. Informed consent ensures that the participants have received a fair explanation of the study and been informed that they can discontinue participation and withdraw the consent at any time.

Demands for informed consent are now included in various national and federal legislations. Official rules are typically set out by national data inspectorates or research ethics committees concerning content and form of such documents. Consent documents are therefore usually comprehensive to cover ethical guidelines and will often require extensive reading. However, the main challenge for people with dyslexia regards reading and word decoding, which typically causes slower reading rates

and/or decoding errors [20]. Consequently, researchers do not know whether the participant has comprehended all the written information before signing, which may cause serious violation of research ethics. Reduced short-term memory may also be a challenge, because participants may quickly forget information they were given at the beginning of a session.

3.3 Method of data collection

In IIR experiments the common procedure is as follows:

- Pre-experiment survey,
- Training for the system,
- Searching system with the given information needs and
- Post-experiment survey.

Most of these procedures, apart from searching the system, are optional depending on the experimental research question but do require thought. The issue raised above in terms of accessibility of documents applies to any survey carried out during the experiment. The key issue is the tasks the users will undertake whilst doing the experiment, and what data will be collected. Users' search behaviour can be logged [23] and eye-tracking can be used [17] in order to collect quantitative data about their activities. Participants must of course be fully appraised of the data being collected about them and why.

The task itself will have a significant impact on the variable used. IIR experiments typically require that at least two topics are used, to ameliorate the topic effect as they can vary in complexity and scope. Experiments often use many more, e.g. the TREC-6 Interactive track [50] used six topics in the experiments. However, presenting too much information to people with dyslexia is problematic (for reasons stated above). Presenting as many as six topics to a user with dyslexia will undoubtedly impact on short term memory capacity and lead to tiredness very quickly. Senior members of the IIR community raised this issue with one author when the idea of experimenting with users with dyslexia was first mooted [31]. For this reason, it is better to use one topic, but this means the topic effect needs to be addressed in a different way (this cannot be ignored due to its importance). This issue also affects the randomization of tasks. If a between subject experiment is carried out then the experimenter needs to ensure that topics are balanced between users in the dyslexia group and the control group. Finally, the experimenter needs to think about the total time for experiments, due to tiredness the user group feels when undertaking intellectually intensive activities.

3.4 Reporting results

There are clear rules both in research ethics and methodology to anonymise data when reporting results. However, there are no guidelines regarding how to refer to participants. There are different practises on how to refer to such user groups, often related to which model of disability that is used, for instance the medical model, social model or the gap-model [18] and terminology often change over time. It is always good practice to anonymise the participants of the study, but often participants in the control group and dyslexia group will be labelled differently so that contrasts can be made in between subject designs.

Another key issue is to keep the participants informed of the results. Although this might be standard procedure, it is particularly important that vulnerable user groups understand that their activities have real worth.

4 REVISED METHODS IN IIR

In this section we suggest some revised methods in IIR to deal with the key issues presented above. According to Newell et al. [36] studies on marginalised groups, for instance people who are elderly or have an impairment, require a different approach than traditional user studies. It is argued that designers need to develop an empathetic relationship with the users rather than treating them as "test subjects". Newell, Gregor, Morgan, Pullin and Macaulay [36] claim that building such a relationship is not always possible in traditional laboratories, where experimenters observe participants from a distance. This approach is referred to as User-Sensitive Inclusive Design [36]. Empathy is an important aspect of all user studies, especially when the test group comprise vulnerable users, and is important in all phases of user studies, from the planning, initial contact with participants and referral of results.

4.1 Sampling

Ethically, it is favourable that researchers are not allowed access to lists of people with impairments. Consequently, it is not possible to use sampling frames or statistical methods to get a random, representative sample of people with dyslexia. The researchers have to rely solely on recruitment by people volunteering, which may cause sample bias. It is therefore important to consider where and how information about the study is given. One solution is for instance cooperation with user organisations, for instance the British Dyslexia Association [7], students disability services at universities or social media such as Facebook groups that are particularly relevant for the target group.

Balancing the participant group according to variables such as gender and age is also important. For instance, dyslexia is more prevalent among males than females [43], and this imbalance should be reflected in the sample. However, in our experiences and as reported by other researchers [26], females are more likely to volunteer for studies than males. As a consequence, there is a risk of an unequal gender distribution, which may be particularly skewed since the sample should consist of a higher portion of males than females.

4.2 Informed consent

Informed consent implies that participants have read, understood and signed comprehensive documents describing the study. However, reading difficulties can cause decoding errors. Moreover, reduced reading speed can make reading recruitment letters a time-consuming and potentially exhausting activity. As a consequence, some participants may sign consent forms

without comprehending the actual content of the document. Conducting a study where participants have signed consent forms without reading and / or fully understanding the content would be a serious violation of research ethics. The experimenter therefore needs to ensure that the participant has comprehended all the information provided.

Difficulties with informed consent have been reported for other vulnerable user groups, such as elderly people with dementia [1] or participants with severe cognitive impairments [8; 38]. Astell, Alm, Gowans, Ellis, Dye and Vaughan [1] suggest that experiments undertaken with users who have dementia should first send out letters about the study to give them time to understand its purpose, then repeat the information later verbally and have the participants consent with a neutral third-party present.

In three of the studies we conducted [3-5], additional verbal information was provided at the start of the experimental session to ensure that the participants had received and comprehended all the necessary information. The use of accessible technologies such as screen-readers can also be used to read out materials, since multimodal communication can assist users with dyslexia in understanding [32]. Another option is to make a short sound recording or movie that can be transmitted to the participants.

Researchers have reported that text format may affect the reading process [12; 32; 44]. Consequently, the researchers have to consider the use of font types, font sizes, spacing between letters and lines and the use of colours, to support the readers with dyslexia to the largest possible extent.

Reduced short-term memory of people with dyslexia is another potential challenge. There is a risk that users may not remember all the information they have been given, for example that they can discontinue participation and withdraw the consent at any time. Consequently, it might be necessary to consider repeating the most important information, especially if the participants are involved in several sessions over time.

4.3 Method of data collection

There are several issues related to methods for data collection. In this section we address diagnoses and the experimental design of topics and tasks.

4.3.1 Diagnosis. Diagnostic papers are highly sensitive, and potential participants may avoid volunteering for studies where they need to share such information with outsiders. A solution to this issue is to rely solely on self-reporting of diagnoses or the researchers conduct screening tests themselves. There are two main issues with self-reporting: does the participant actually have a valid diagnosis, and if control users are included, are the experimenters sure that none of these participants have dyslexia? The first issue might be especially important if the participants are given remuneration, since there is a risk of people volunteering simply to receive the payment. Further, people may think they have a diagnosis, but without proper evaluation it is not possible to ensure that the right diagnosis has been set.

The second issue with self-reporting of dyslexia concerns the control users. It is crucial to ensure that the control participants do not have any degree of dyslexia, since the presence or absence of this diagnosis is typically the controlled variable. It is reported that people may not be aware of their dyslexia, and undiagnosed adults is not uncommon [51]. Since people may not be aware of their dyslexia diagnosis, there is a need for screening tests of the control group as well as the self-reported users with dyslexia. In one of our studies [2] several control users had to be excluded due to low scores on the screening tests, confirming the need for screening all participants.

Another concern is how to deal with the test scores of control users. According to International Test Commission [22], researchers should provide users with feedback of test results. We argue that this is especially important if the test scores for control users are indicative of dyslexia, where the participants may need further diagnostic tests conducted by professionals to receive a proper diagnosis. For example, with student participants the researcher should know the relevant student support services and direct the participant to the relevant person in that service.

One quick way to test whether a person might have dyslexia or not is to use the Adult Dyslexia Checklist developed by the British Dyslexia Association [46]. This screening test does not provide enough information for diagnostic assessment, but may indicate assessment needs. The test contains 15 questions related to typical issues associated with dyslexia, for example whether the person confuses words, have difficulties reading out loud and have trouble telling left from right. This test is no substitution for a diagnosis from a qualified psychologist, but might be a helpful screening tool.

Word Chain tests are also widely used in research [11], and is based on word recognition skills. The procedure is as follows: the participant is presented with a certain amount of word chains, where several words are put together without blank space. The task is to divide the word chain into the proper words by marking the white space, and the user is given a certain amount of time to complete the task. Points are given based on the number of correct word chains completed on time, and normative data exists for different age ranges. These tests are quick to complete, easy to evaluate and are developed in several languages and variations [11]. However, this test must also be regarded as a screening tool, and not a diagnostic test.

Dual diagnosis represents another challenge. It might be difficult to investigate the effect of dyslexia on information searching in contexts with a presence of other diagnoses such as ADHD or dyscalculia. Depending on the research aim, it may be necessary to remove such participants from the study results. Moreover, dyslexia occurs in different degrees and with variations. Consequently, a thorough psychological evaluation of each participant may be necessary to have a proper understanding of the cognitive profile of each user. Examples of characteristics to test are decoding skills, concentration skills, reduced short-term memory capacity and rapid naming skills, all of which may affect the ability to search for information.

There are a number of relevant tests available that are relatively easy to conduct [19; 48; 53]. A widely used test is the open source Victoria Stroop test, that may provide data on concentration and rapid naming skills [48]. Memory capacity tests may also be useful, for instance a Digit span test [52] or a Corsi Block-tapping test [48]. Both of these memory tests are freely available in the open source software "PEBL: The Psychology Experimental Building Language" [34].

Overall, it is important to only include tests that are necessary, explain to the participants why they are conducted and to complete them according to ethical guidelines. When working with vulnerable user groups, it is especially important to reduce the test-takers anxiety and avoid creating unnecessary anxiety [22].

Another general issue is that some researchers combine users with very different diagnoses in one study, for instance autism, dyslexia and ADHD without discussing the potential problems concerning the inclusion of users with needs that differ substantially or are conflicting. We recommend focusing on one diagnosis at a time whenever possible. Otherwise, it should be clearly stated and discussed that several diagnoses are studied, and results should be isolated for each user group and compared to see whether there are differences according to diagnosis.

4.3.2 Experimental tasks. As stated in Section 3.3, care must be taken to ensure participants are treated ethically and fairly e.g. not tiring them out. Firstly, the users should not be involved in long sessions with many topics. In MacFarlane, Al-Wabil, Marshall, Albrair, Jones and Zaphiris [28] and MacFarlane, Albrair, Marshall and Buchanan [29] experiments were restricted to around 1 hour, with actual searching time approximately 30 minutes. Two topics can be used in the experiment, but each user only gets one. The topics are split in both the dyslexia and control group, but both get the same topics (see Table 1). Topics should be split amongst each type of user as evenly as possible (e.g. round robin, so that users do not all start with the same topic), and the balance between types of users should nearly be equal (e.g. C should be around the same size as D).

Randomization of tasks is not always possible. In Berget and Sandnes [4] and Berget and Sandnes [5] users were asked to solve search tasks with various degrees of difficulty. The tasks were formulated quite specifically, often including relevant search terms, because the purpose was to test how reduced spelling skills caused by dyslexia may affect query formulation. The study was divided into two sessions conducted some months apart, with searching times of approximately 20 minutes for each session. Each participant solved the same 10 tasks, but were given a different set for the two sessions. The varying difficulty levels of the tasks combined with the vulnerability of participants with dyslexia made it impossible to randomize the tasks. Consequently, all participants solved the tasks in the same order, starting with a very simple task to provide a sense of coping. This was followed by tasks that were increasingly difficult. An easier task was given in the middle of the session with the purpose to encourage the participants, while the most difficult task was solved last. Several participants commented after the experiment that if the most difficult task had been presented first, they would have lost confidence and terminated the session. These attitudes confirm the need to consider the vulnerability and feeling of coping of the participants over rigid methodological procedures concerning randomization of tasks. However, the researcher needs to be aware of the consequences of this methodological choice.

Table 1: Topic distribution (C=Control users, D=Users with Dyslexia)

User Type	Topic A	Topic B
Control	C/2	C/2
Dyslexia	D/2	D/2

4.4 Reporting results

Here we report best practice when dealing with the data collected in experiments involving users with dyslexia.

4.4.1 Diagnosis. Experimenters should be clear about what (if any) procedure they used to ensure the controls and users with dyslexia are distinct groups, their strategy with dealing with participants with undiagnosed dyslexia and how they deal with them in the study. Given that dyslexia describes a spectrum of cognitive abilities, the researcher should provide an overview of the profile of the cohort and provide some idea as to how the participants match the overall population of people with dyslexia. In cases where the difference between various types of participants with dyslexia is examined (e.g. severe vs mild forms), then the cognitive profile differences need to be clearly explained and again the research needs to justify the placement of users into the particular groups. In such cases, the screening tests described in Section 4.3.1 can be useful to get an indication of the severity of the dyslexia for each person. Diagnosis can also be used when examining experimental results. Moreover, dual diagnoses need to be addressed, as discussed in Section 4.3.1, and preferably statistical tests should be run with and without dual diagnosis participants (and their controls if applicable).

4.4.2 Experimental results. Variables collected when examining the cognitive profile can be used in statistical analysis with search variables (e.g. such as number of iterations in a session, number of terms per query or time spent per query) and correlations can be particularly useful. For example MacFarlane, Albrair, Marshall and Buchanan [29] took one variable (documents judged irrelevant) and did correlations with the cognitive variables. In this case Digit Span (a measure of short-term memory capacity) showed a strong correlation with that search variable. We concentrated on this search variable as it was the only one in which the difference between the dyslexia group and the control group was significant.

Our advice would be to identify an initial variable or variables that have demonstrated significant difference and focus on these. However, what must be done is to ensure that there is a real difference between the two groups, and there is no topic effect i.e. topic effects are considered to be important in the field and variables have to be cross checked accordingly. The design of the experiment outlined in section 4.3.1 is therefore critical, and an orthogonal topic comparison is necessary (as per Table 1).

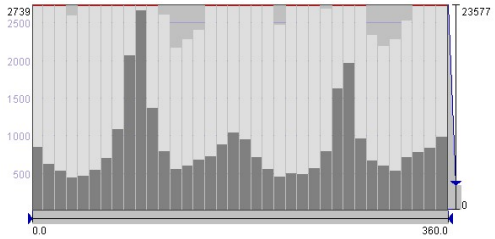

Figure 1 – Control group frequency distribution of eye movements

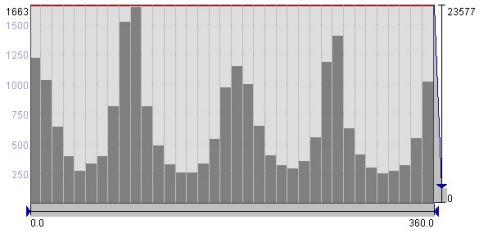

Figure 2 – Dyslexia group frequency distribution of eye movements

In MacFarlane, Albrair, Marshall and Buchanan [29] the only statistically significant variable showed no topic effect i.e. there was no significant difference between the topics. Thus, if no topic effect is demonstrated on a given variable (documents judged irrelevant), then any difference from other comparisons is real i.e. the difference between controls and users with dyslexia on that variable. This is the key way to ensure that users are treated ethically and some rigour in the experimental results is adhered to.

Comparing search variables between the two user groups may also be purposeful. In the context of search variables, users with dyslexia and the control group can be compared with for example t-tests to get a better understanding of which parts of the search process that are most affected by the impairment. These types of comparisons may be useful to apply to different types of search systems. For instance, Berget and Sandnes [4], found that there were significant differences between users with dyslexia and control users in misspellings, search times and number of queries needed to solve a task in a system with no query-building aids and a high demand for correct spelling. In another experiment using a system with query-building aids and a higher tolerance for misspellings, Berget and Sandnes [5] found that the differences in spelling skills did not affect the overall performance due to a higher tolerance for errors implemented in the system. Consequently, analysing several variables and including different types of systems may demonstrate the potential in UI design to counteract or entirely remove the potentially negative effect of an impairment on search performance.

Methodological choices are also, as in any experiment, important. Regarding dyslexia, eye-tracking has proven especially useful, since the eye-movements may reveal certain search behaviours that are not directly visible in variables such as search times or misspellings. For instance, MacFarlane, Buchanan, Al-Wabil, Andrienko and Andrienko [30] found that people with dyslexia backtracked when viewing results and documents during search, which is probably caused by short-term memory challenges (forgetting and looking back as a reminder).

Using visualisations (see Figure 1 and 2) can provide very useful contrasts on different users' interaction behaviours. For instance, a comparison of figures 1 and 2 show change of eye direction on the UI, compiled in a histogram from 0^0 to 360^0. In these examples eye movements at $0^0/360^0$ and 180^0 are more frequent indicating more vertical eye movements in users with dyslexia than control users. In a study by Berget and Sandnes [5], eye data revealed that people with dyslexia did not look at the suggestions provided by the autocomplete function as extensively as the control users, due to an intense focus on the keyboard during query input.

4.4.3 Referring to users. The strategy for referring to people with impairments have varied over time, and has among other factors been closely related to the dominating model on disability. According to Grue [18], an especially discussed topic has been the impairment/disability dichotomy. It is now a common view that a person has an impairment, not a disability. In contrast, a disability is something that occurs when there is a mismatch between the demands from the society and the person's abilities. Moreover, a basic rule seems to be referring to people first, and not the disability, for instance "person with a cognitive impairment", not "cognitively disabled" or "cognitively impaired". In this case we would recommend using phrases such as 'person with dyslexia'.

Another issue is how one refers to the control group. In some contexts, the words neurotypical is used for the control users, in other cases one simply uses "control group" or "participants without dyslexia". It is important to avoid using "normal" as a measurement, thus implying that the participants who diverge from the "standard" are "not normal". It is also important to be aware of terminology changes over time. For instance, the phrase "word blindness" has been replaced by "dyslexia" [45], and should not be used, since it may give a wrong impression of the aetiology of the diagnosis. Consequently, when referring to older studies researchers cannot always uncritically adopt the same terminology applied there, they must ensure that they use current terminology from recent literature.

5 SUMMARY OF RECOMMENDATIONS

Our stated aim was to provide some ideas for best practice in conducting IIR experiments involving users with dyslexia, and the many issues with the standard form of experimentation that researchers must consider if they are to undertake the work ethically. The key issue is that these users are regarded as being a vulnerable group, due to low self-esteem, tendency to tiredness and anxiety. The standard IIR experimental method must be adapted to ensure that ethical considerations are adhered to, but also ensuring that the methods used are rigorous and the results produced are useful and transferable to the IIR community either for undertaking further experimentation with the group or to inform system design, or both. We have not addressed the issue of naturalistic studies, but many of the ideas here in terms of sampling and informed consent do also apply to such contexts.

User studies on various impairments have revealed important design issues in search user interfaces, and topics that need to be addressed by the research community. In accordance with the universal design perspective, several of these issues would be problematic for other users as well, and measures to make the systems more accessible would also be beneficial for people without impairments.

We summarise our recommendations based on our prior experience outlined above in the following sub-sections with a particular focus on undertaking research involving users with dyslexia (we assume the standard IIR experimental procedure).

5.1 Experimental Design

When considering study design there are a number of key issues around users and topics that need to be considered. Researchers should be clear about how they will distinguish between the users with dyslexia and the control group, and if necessary, between different types of dyslexia, e.g. mild vs severe forms. Given the ethical issues with requesting prior diagnosis, we strongly recommend that researchers consider the use of a wide variety of psychological tests which are freely available [34; 46; 48]. Researchers should put in place strategies to deal with undiagnosed users with dyslexia in the control group i.e. where to get advice and who to contact for a full diagnostic evaluation. The user should be provided with a full profile of their diagnostic test results if required. This is a significant ethical issue, which researchers must think carefully about when for example completing ethics panel applications. Other variables such as the gender balance should be considered given the prevalence of dyslexia in the male population.

Once the researchers have a clear focus on the type of user with dyslexia they will work with, various aspects of the study design need to be considered carefully. In terms of topics and/or tasks, a limited number should be selected, but should not be restricted to just one type of task or one topic – the task/topic effect is too profound. At least two topics should be chosen, but should be balanced out to ensure that equal number of topics are distributed to users in both the dyslexia group and the control group (see Table 1). If necessary, distribute the topics within each group based on gender. Researchers need to think about the complexity of the topics and its potential effect on for instance tiredness and self-esteem of users with dyslexia. If tasks rather than topics are used, a limited number should be utilized, and the researcher should think about the balance between complexity and number (many simple tasks vs. fewer complex tasks). All sub groups should be as balanced as possible within the sampling constraints.

Next the researcher needs to think about the materials to be used in the experiment. What type of survey instruments will be used e.g. questionnaires, interviews, observations? Logging of the searches may also be considered [28; 29]. How will informed consent be obtained from the user, and how will all materials be presented in an accessible way?

Researchers should not only think about the form of information sheets, consent forms and questionnaires (e.g. clear wording without excessive text) but the type of accessible technology that may be required by the users (e.g. screen readers and colour schemes on the user interface). Before proceeding, ethical approval for the study must be obtained from an appropriate panel.

5.2 Experimental conduct

Once the experiment has been designed and appropriate ethical approval has been obtained the work can be carried out and participants recruited for the study. Informed consent must be given by the user before the experiment is started. At any stage in the experiment, the researcher should be prepared to step in and give the user appropriate assistance (e.g. with assistive technologies). Also, the researcher should keep in mind that the participant might have forgotten information given at the beginning of the experiment, and consider repeating the most fundamental issues, especially if the experiment is conducted over more than one session.

Before starting the experiment properly, it is important to carry out any diagnostic process beforehand. All users must be subject to the same tests (useful for analysis later on), the type depending on the aims of the study. If control users are recruited, and found to have a similar profile to users with dyslexia, provide them with advice on what to do next together with the data collected so that a professional therapist can provide them with the appropriate help. In terms of the experiment itself, we strongly recommend that they are removed from the study and not placed in the dyslexia group. Users with dyslexia have had a full diagnostic assessment, whereas undiagnosed participants do not, and putting them into that group introduces uncertainty, impacting on the reliability of the study. However, the undiagnosed user should be allowed to complete the experiment for their own interest, and to obtain the appropriate reward (in having volunteered for the study in the first place).

Once the diagnostic test has been carried out the experiment can be done in earnest. This may well start with a pre-experiment survey, collecting demographic information e.g. gender, age and perhaps the level of prior experience with the type of search system being investigated. Accessible technology should be used where necessary e.g. screen readers for

questionnaires. After this a short training session should take place to familiarise the user with the system, either with a test topic or with the users' own information need. This should be long enough for the user to learn how to use the system, but should not over burden the user cognitively (taking into account the work they will undertake in the main experiment).

The main experiment should then take place. We recommend restricting the experiment to no more than one hour, with around 30 or so minutes devoted to the main searching task. Researchers should think clearly about the balance between the different elements of the IIR process. Researchers should be ready to step in when participants are having problems, and if it is clear that the users struggle due to the system itself (e.g. the UI causes visual problems), then the researcher should be prepared to either make changes or in extreme cases stop the experiment. In such cases user's data should be omitted from the study.

When the experiment is completed, any post experiment survey can be carried out, using the same strategy and methods as with any pre-experiment survey. The researcher should take time to wrap up and reassure the participant if and when required. It may be appropriate to return to the issue of informed consent to ensure the user fully understood the purpose and rationale of research carried out. For participants who have struggled and showed frustration and/or anxiety about the performance, the researcher should take the time to explain that the study was not about performing at a certain level, but to rather understand how the UI design can be changed to better accommodate their search behaviour. Consequently, the participant will hopefully feel less frustrated and that the effort was purposeful.

5.3 Reporting and dissemination of results

A key issue with the analysis stage is putting thought into the analysis of variables collected. The first thing to do is to examine the cognitive variables and undertake a statistical analysis on group membership – in particular to ensure that control participants are not undiagnosed users with dyslexia. Typically, comparisons between participants with dyslexia and the control users or users with severe and mild dyslexia will be carried out, both on the cognitive variables and the search variables. Correlations between cognitive and search variables can be usefully carried out. Any demographic information e.g. gender can also be used as an extra dimension of analysis.

One key issue must be addressed in the analysis – that of task or topic. For example, the variables used to examine the differences between users with dyslexia and the controls can also be subject to a topic comparison (as per Table 1). The class of tasks can be treated in a similar way. If any significant difference is found on a given variable on the topic, then there is a clear topic effect to factor into the analysis. For example, in cases where a topic effect is demonstrated on a variable, statistical significance on comparisons between users in the dyslexia group and the control group on that variable need to be treated with care – i.e. the effect of dyslexia is not conclusively demonstrated. Further analysis may be necessary to make conclusive statements – one strategy is to reject any significant differences on the cognitive variable if a topic effect is shown [28; 29]. A further method which can prove useful is the application of multiple regression analysis [13], which can be used to demonstrate which variables are the most significant. Researchers should consider different visualisation techniques for the display of aggregate data, for example the eye tracking examples provided in figures 1 and 2. More examples can be found in MacFarlane, Buchanan, Al-Wabil, Andrienko and Andrienko [30].

If literals have been collected in the pre and post experiment survey, then various forms of qualitative analysis can be carried out for example discourse analysis. Mixed method approaches can be a very useful strategy in the domain. We argue that any statistical analysis of the user behaviour and cognitive profile can be enhanced by the examination of user views of the search experience, picking up issues on usability and accessibility that cannot be detected by quantitative methods alone.

Finally, when writing up the results, researchers should think clearly about how they refer to their participants – particularly given that the users recruited may wish to see the published article.

6 CONCLUSION

In this paper we argue for the use of a revised IIR approach, that builds on the knowledge already gained in the area over many years. In our experience this has yielded positive results, and we have been able to make contributions to knowledge through the methodology. A question to be raised is whether this methodology is generalizable to other types of impairments. In the authors view, this needs to be considered on a case by case basis – it is entirely possible that some impairments put serious constraints on the IIR methodology making it unusable, in which case a different approach is required. We hope this paper is the start of a serious examination of the issue of impairments and methods in the IIR community.

REFERENCES

[1] A. Astell, N. Alm, G. Gowans, M. Ellis, R. Dye and P. Vaughan. 2008. Involving older people with dementia and their carers in designing computer based support systems: Some methodological considerations. *Universal Access in the Information Society*, 8(1), 49.

[2] G. Berget. 2016. *Search and find?: An accessibility study of dyslexia and information retrieval*. University of Oslo, Oslo, Norway.

[3] G. Berget, F. Mulvey and F. E. Sandnes. 2016. Is visual content in textual search interfaces beneficial to dyslexic users? *International Journal of Human-Computer Studies, 92-93*, 17-29.

[4] G. Berget and F. E. Sandnes. 2015. Searching databases without query-building aids: Implications for dyslexic users. *Information Research, 20*(4), paper 689.

[5] G. Berget and F. E. Sandnes. 2016. Do autocomplete functions reduce the impact of dyslexia on information-searching behavior?: The case of Google. *Journal of the Association for Information Science and Technology, 67*(10), 2320-2328.

[6] P. Borlund. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90.

[7] British Dyslexia Association, 2018. *British Dyslexia Association*. Retrieved from https://www.bdadyslexia.org.uk/.

[8] L. Cameron and J. Murphy. 2007. Obtaining consent to participate in research: The issues involved in including people with a range of learning and communication disabilities. *British Journal of Learning Disabilities, 35*(2), 113-120.

[9] J. M. Caroll. and J. E. Iles, 2006. An assessment of anxiety levels in dyslexic students in higher education. *British Journal of Educational Psychology, 76*(3), 651-662.

[10] City University of London, 2018. *Research ethics City University, London.* Retrieved from https://www.city.ac.uk/department-computer-science/research-ethics.

[11] A. E. Dahle and A.-M. Knivsberg, 2014. Internalizing, externalizing and attention problems in dyslexia. *Scandinavian Journal of Disability Research, 16*(2), 179-193.

[12] L. Evett and D. Brown, 2005. Text formats and web design for visually impaired and dyslexic readers: Clear text for all. *Interacting with Computers, 17*(4), 453-472.

[13] A. Fourney, M. R. Morris, A. Ali and L. Vonessen, 2018. Assessing the readability of web search results for searchers with dyslexia. In *Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, New York, NY, USA, 1069-1072.

[14] C. Frankfort-Nachmias, D. Nachmias and J. Dewaard, 2015. *Research methods in the social sciences*. Worth, New York, NY, USA.

[15] E. Germanò, A. Gagliano and P. Curatolo, 2010. Comorbidity of ADHD and dyslexia. *Developmental Neuropsychology, 35*(5), 475-493.

[16] J. Gibbs, J. Appleton and R. Appleton, 2007. Dyspraxia or developmental coordination disorder?: Unravelling the enigma. *Archives of Disease in Childhood, 92*(6), 534-539.

[17] J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott and A. M. Wichansky, 2002. Eye tracking in web search tasks: Design implications. In *Proceedings of the Proceedings of the 2002 symposium on Eye tracking research & applications, New Orleans, Louisiana 2002*. ACM, New York, NY, USA, 51-58.

[18] J. Grue, 2011. Discourse analysis and disability: Some topics and issues. *Discourse & Society, 22*(5), 532-546.

[19] J. Hatcher, M. J. Snowling and Y. M. Griffiths, 2002. Cognitive assessment of dyslexic students in higher education. *British Journal of Educational Psychology, 72*(1), 119-133.

[20] M. Hebert, X. Zhang and W. Parrila, 2018. Examining reading comprehension text and question answering time differences in university students with and without a history of reading difficulties. *Annals of Dyslexia, 68*(1), 15-24.

[21] H. Hill, 2013. Disability and accessibility in the library and information science literature: A content analysis. *Library & Information Science Research, 35*(2), 137-142.

[22] International Test Commission, 2001. International Guidelines for Test Use. *International Journal of Testing, 1*(2), 93.

[23] B. J. Jansen, 2006. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*(3), 407-432.

[24] S. Jeffries and J. Everatt, 2004. Working memory: Its role in dyslexia and other specific learning difficulties. *Dyslexia*, 10(3), 196-214.

[25] D. Kelly, 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval, 3*(1–2), 1-224.

[26] S. E. Kelly, T. D. Spector, L. F. Cherkas, B. Prainsack, and J. M. Harris, 2015. Evaluating the consent preferences of UK research volunteers for genetic and clinical studies. *PLOS ONE, 10*(3), e0118027.

[27] K. Landerl, B. Fussenegger, K. Moll, and E. Willburger, 2009. Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology, 103*(3), 309-324.

[28] A. MacFarlane, A. Al-Wabil, C. R. Marshall, A. Albrair, S. A. Jones and P. Zaphiris, 2010. The effect of dyslexia on information retrieval: A pilot study. *Journal of Documentation, 66*(3), 307-326.

[29] A. MacFarlane, A. Albrair, C. R. Marshall and G. Buchanan, 2012. Phonological working memory impacts on information searching: An investigation of dyslexia. In *Proceedings of the Proceedings of the 4th Information Interaction in Context Symposium, Nijmegen, The Netherlands, 2012*. ACM, New York, NY, USA, 27-34.

[30] A. MacFarlane, G. Buchanan, A. Al-Wabil, G. Andrienko and N. Andrienko, 2017. Visual analysis of dyslexia on search. In *Proceedings of the Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, Oslo, Norway, 2017*. ACM, New York, NY, USA, 285-288.

[31] A. MacFarlane, H., Petrie and S. A. Jones, 2005. A user study on the effect of dyslexia on information retrieval. In *Proceedings of COLIS 2005 Workshop on Evaluating User Studies in Information Access*, 31-38.

[32] J. E. McCarthy and S. J. Swierenga, 2010. What we know about dyslexia and Web accessibility: A research review. *Universal Access in the Information Society, 9*(2), 147-152.

[33] T. Mortimore and W. R. Crozier, 2006. Dyslexia and difficulties with study skills in higher education. *Studies in Higher Education, 31*(2), 235-251.

[34] S. T. Mueller, and B. J. Piper, 2014. The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods, 222*, 250-259.

[35] B. A. Nalavany, L. W. Carawan, and L. J. Brown, 2011. Considering the role of traditional and specialist schools: Do school experiences impact the emotional well-being and self-esteem of adults with dyslexia? *British Journal of Special Education, 38*(4), 191-200.

[36] A. F. Newell, P. Gregor, M. Morgan, G. Pullin and C. Macaulay, 2011. User-Sensitive Inclusive Design. *Universal Access in the Information Society, 10*(3), 235-243.

[37] R. I. Nicolson and A. J. Fawcett, 2011. Dyslexia, dysgraphia, procedural learning and the cerebellum. *Cortex, 47*(1), 117-127.

[38] M. Nind, 2008. *Conducting qualitative research with people with learning, communication and other disabilities: Methodological challenges*. National Centre for Research Methods, University of Southampton, Southampton, United Kingdom.

[39] E. S. Norton and M. Wolf, 2012. Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology, 63*(1), 427-452.

[40] S. Novita, 2016. Secondary symptoms of dyslexia: A comparison of self-esteem and anxiety profiles of children with and without dyslexia. *European Journal of Special Needs Education, 31*(2), 279-288.

[41] T. M. Perez, M. Poncelet, E. Salmon and S. Majerus, 2015. Functional alterations in order short-term memory networks in adults with dyslexia. *Developmental Neuropsychology, 40*(7-8), 407-429.

[42] R. L. Peterson and B.F. Pennington, 2015. Developmental Dyslexia. *Annual Review of Clinical Psychology, 11*(1), 283-307.

[43] J. M. Quinn and R. K. Wagner, 2015. Gender differences in reading impairment and in the identification of impaired readers: Results from a large-scale study of at-risk readers. *Journal of Learning Disabilities, 48*(4), 433-445.

[44] L. Rello and R. Baeza-Yates, 2017. How to present more readable text for people with dyslexia. *Universal Access in the Information Society, 16*(1), 29-49.

[45] T. A. Serry and L. Hammond, 2015. What's in a word? Australian experts' knowledge, views and experiences using the term dyslexia. *Australian Journal of Learning Difficulties, 20*(2), 143-161.

[46] I. Smythe and J. Everatt, 2001. *Adult checklist*. Retrieved from https://www.bdadyslexia.org.uk/common/ckeditor/filemanager/userfiles/Adult-Checklist.pdf.

[47] M. F. Story, 1998. Maximizing usability: The principles of universal design. *Assistive Technology, 10*(1), 4-12.

[48] E. Strauss, E. M. S. Sherman and O. Spreen, 2006. *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford Press, Oxford, United Kingdom.

[49] M. M. Terras, L. C. Thompson and H. Minnis, 2009. Dyslexia and psycho-social functioning: An exploratory study of the role of self-esteem and understanding. *Dyslexia, 15*(4), 304-327.

[50] Text Retrieval Conference, n.d. *TREC-6 interactive track specifications*. Retrieved from https://trec.nist.gov/data/t6i/trec6spec

[51] M. Warmington, S. E. Stothard and M. J. Snowling, 2013. Assessing dyslexia in higher education: The York adult assessment battery-revised. *Journal of Research in Special Educational Needs, 13*(1), 48-56.

[52] D. Wechsler, 1997. *Wechsler Adult Intelligence Test*. The Psychological Corporation, San Antonio, TX, USA.

[53] G. S. Wilkinson and G. J. Wilkinson, 2006. *Wide Range Achievement Test*. Psychological Assessment Resources, Florida, USA.