

MAUU5900
MASTER'S THESIS
in
Universal Design of ICT

May 2018

**Simplifying and Improving Effectiveness of Image Description
for Accessibility using Sample Cues**

Dhruba Dahal

Department of Computer Science
Faculty of Technology, Art, and Design



Oslo Metropolitan University
Norway

Preface

This master's thesis will report on the use of sample cues in order to help Web users write more accessible image descriptions while posting images on Web. This study will evaluate the degree of accessibility of an image description based on the National Center for Accessible Media (NCAM) guidelines. The research titled 'Image Accessibility on Web' was proposed by OsloMet University in order to address problem of increment of inaccessible images on Web. Through this research, I gained in-depth knowledge about how one can conduct an online experiment in order to explore the area of Human-Computer Interaction (HCI) and especially, the concept of Universal Design of ICTs in terms of image accessibility became clearer to me.

There are several people who helped me very well during this research. I am thankful to my supervisor, Associate Professor Raju Shrestha, for the guidance, support, and useful tips along the way. I would also like to thank my co-supervisor, Asieh Abolpour Mofrad, for her support, interest, and enthusiasm in my research. I cannot remain without giving thanks to my wife, lechha Koirala, who managed everyday home activities alone including our newly born son, Dikchhyant Dahal. Due to her support, I got enough time to engage with my study.

Third, I would like to thank my participants who gave their valuable time for the experiment with great interest and enthusiasm. It otherwise would not be possible to come up with these results without them. I would also like to thank my friend Sandeep who gave me valuable insights regarding image accessibility during the initial phase of this research.

Lastly, I would like to dedicate this master's thesis to my parents, Oum Nath Dahal and Gita Dahal, who always remind me to work hard and believe on my ability to achieve something that is good for everyone.

Dhruba Dahal

Oslo, May 18, 2018.

Summary

Universal design (UD) of information communication technology (ICT) is a basic principle that confirms the accessibility of ICT products and services to avoid discrimination and to support the equal ability participation in the society to enhance democracy. An image is a non-textual Web content which demands an equivalent image description to be accessible and usable for the users who have different abilities.

Previous research suggests that the significant number of images exist on the Web which do not contain accessible and usable text descriptions. Literature also suggests that describing an image is a complex task for the Web workers and normal users. To the knowledge of this study, there is almost no research available in image accessibility which incorporates different types of images.

This research's aim is to fulfill the existing gap of not having the proper solution to help Web users in writing usable image descriptions. This study investigated the effectiveness of providing sample cues in making image description comply with the set of guidelines suggested by National Center for Accessible Media (NCAM) through an online experiment.

Three different types of images: graph, map, and general image are provided in three different steps with no cue, random cue, and similar cue respectively. Text descriptions, Likert-type ratings on the compliment to NCAM guidelines, and time data are collected from two separate groups—participants and the judgement group. The experimental result is analyzed using several statistical tests.

The results from this research suggest that providing the similar sample cue is effective writing an accessible image description in case of all the three types of images. Similar sample cues make it easier to write a description which follows the NCAM guidelines. The random sample cues are better than not having any cue. However, there is no significant difference in time it takes to write image descriptions in the three cases.

Table of Contents

Preface.....	i
Summary.....	ii
List of Figures.....	vii
List of Tables.....	ix
List of Acronyms.....	x
Terms and Definitions	xi
1. Introduction	1
1.1 The Statement of the Problem	2
1.2 Study aims	3
1.3 Research Questions.....	3
1.4 Thesis Organization	3
2. Background	4
2.1 The Concept of Universal Design.....	4
2.1.1 Web Accessibility.	5
2.1.2 Image Accessibility.....	6
2.2 Social Aspects of Image Accessibility	8
2.2.1 Information Society.....	8
2.2.2 Digital Divide.	8
2.2.3 Diverse Content Authors on Web.....	9
2.3 Technical Aspects of Image Accessibility	10
2.3.1 Web Authoring Tools.....	10
2.3.2 Content Management Systems and Accessibility.....	13
2.3.3 Assistive Technology.....	14

2.4 Guidelines for Accessibility.....	15
2.4.1 Web Content Accessibility Guidelines (WCAG).....	15
2.4.2 ATAG.	17
2.4.3 NCAM Guidelines.....	17
2.5 Describing Image	18
2.5.1 Importance of Image Description.....	18
2.5.2 Art of Writing Alternative text.....	19
2.5.3 Image Description Implementation.	20
2.6 Related Works on Image Accessibility	20
2.6.1 Human Powered Authoring Tools for Image Description.....	21
2.6.2 Computer Algorithm Based Image Accessibility tool.	22
2.6.3 Proposed Cue-based Image Description Writing.....	25
3. Cue-based Image Description Writing and Research Methodology	26
3.1 Cue-based Image Description Writing.....	26
3.1.1 Sample cue.	26
3.1.2 Providing Sample cues on User Interface.	26
3.1.3 Feasibility of Providing Similar cues.....	27
3.2 Research Methodology	27
3.2.1 Experimental Research.	28
3.2.2 Research Hypothesis.	30
3.2.3 Research Variables.	31
3.2.4 Within-Group Design	32
3.2.5 Quantitative data.	32
3.2.6 Data Management.....	32
3.2.7 Image type Selection.....	32

3.2.8 Selection of the Guidelines.....	36
3.2.9 Image Description Evaluation Criteria.	37
3.2.10 Statistical Analysis.....	40
3.2.11 Participants.....	42
3.3 Pilot Study Before Actual Experiment.....	44
4. Experiment and Results	46
4.1 Experimental Setup and Approach.....	46
4.1.2 Assignment.	49
4.1.3 Conditions/Steps.	49
4.1.4 Tool for the Experiment	49
4.2 Experimental Results	52
4.2.1 Results with no cue, Random cue and Similar cue for the Common Guidelines..	53
4.2.2 Comparison of Effects of Similar and Random cues for Common Guidelines.....	56
4.2.3 Results with no cue, Random cue, and Similar cue for the Specific Guidelines..	59
4.2.4 Comparison of Effects of Similar and Random cues for the Specific Guidelines..	61
4.2.5 Results with no cue, Random cue, and Similar cue for Overall NCAM Guidelines.	63
4.2.6 Results with the Similar and Random cues Based on Image Types.	65
4.2.7 Results for the Level of Difficulties.	68
4.2.8 Results on Time Taken to Write Image Descriptions.....	70
5. Discussion	71
5.1 Revisiting the aims of the Research	71
5.2 Major Differences and Similarities with Previous Research	73
5.3 Usefulness of the Study Results.....	74
5.4 The Strong and weak Aspects of the Research	75

5.5 Significance of Results	77
5.6 Generalizing the Findings.....	77
5.7 The Potential Criticisms of the Research	78
5.8 Ethical Consideration	78
6. Conclusions & Future work.....	80
6.1 Conclusions from the Research	80
6.2 Future work	80
References.....	82
Appendixes	89
Appendix A: analysis reports.....	89
Appendix B: image description examples written by the participants	95
Appendix C: relational database used in this study	97

List of Figures

Figure 2.1. A graphical user interface to upload a post with an image on Facebook. ...	11
Figure 2.2. A graphical user interface to upload a post with an image on Twitter.	11
Figure 2.3. A graphical user interface for inserting image to a page as a content in WordPress.....	12
Figure 2.4. A GUI for uploading an image on Tumblr blog.	13
Figure 2.5. Related components of Web accessibility (W3C, 2005).....	20
Figure 3.1. A simple user interface having random sample cue while posting a scatter plot graph.	27
Figure 3.2. An example webpage showing human judgement framework for image description used in this study.	39
Figure 3.3. Several participants according to their nationalities.	43
Figure 3.4. Several participants according to their professions.	43
Figure 4.1. An online consent form.	47
Figure 4.2. A Web form to enter participant's information.	48
Figure 4.3. Main page after logging in.	50
Figure 4.4. Second step page with a random example.	51
Figure 4.5. Third step page with a similar example.	52
Figure 4.6. Graph showing the number of image descriptions in compliance with the common guidelines with no cue, random cue, and similar cue for all the three types of images.	53
Figure 4.7. A line graph showing the average effect of sample cues based on the common guidelines for the graph, map, and general image.....	56
Figure 4.8. An analysis report based on the Friedman test for the effect of random and similar cues in compliance with the common guidelines.	57
Figure 4.9. Bar graphs showing the descriptions in compliance with the specific guidelines for the graph while having no, random, and similar cues.	59
Figure 4.10. Bar graphs showing the number of descriptions in compliance with the specific guidelines for the map while having no, random, and similar cues.	60

Figure 4.11. Bar graphs showing the number of descriptions in compliance with the specific guidelines for general image while having no, random, and similar cues.....	61
Figure 4.12. A line graph showing the average effect of example cues based on the specific guidelines for the graph, map, and general image.....	61
Figure 4.13. A line graph showing the number of descriptions in compliance with overall guidelines while having no, random, and similar cues.....	63
Figure 4.14. A line graph showing the number of descriptions in compliance with the overall guidelines based on the image types while having the random and similar cues.	65
Figure 4.15. A line graph showing the level of difficulties experienced by the participants while writing the image descriptions in three different conditions.	68
Figure 4.16. Line graph showing time taken by the participants to write the image descriptions with different sample cues.	70

List of Tables

Table 3.1. Selected images with random and similar images..... 36

Table 4.1. Rank table for similar-random cue combinations for the common guidelines.
..... 58

Table 4.2. Test statistics table for random-no, similar-no, and similar-random cue
combinations for the common guidelines. 59

List of Acronyms

NCAM	National Center for Accessible Media
ICT	Information Communication Technology
UD	Universal Design
STEM	Science Technology Engineering and Mathematics

Terms and Definitions

Accessible Image Description: text description to an image which is reachable, understandable, and equivalent to the information an image is intended to show.

Non-text Contents: Web contents that cannot be read by the screen reader software.

Sample cue: example images having text descriptions that comply with the NCAM guidelines.

General Image: natural photos of living and non-living objects.

1. Introduction

“The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect.”- Tim Berners-Lee (Henry & McGee, 2016).

The above-mentioned statement reflects the significance of Web accessibility. It is about the fundamental design of Web for all people regardless of their hardware, software, language, culture, location, and physical or mental ability and the fulfilment of this goal results in accessible Web with a diverse range of sight, hearing, movement, and cognitive ability (Henry, 2005). The UN Convention—on the rights of people with disabilities— together with the growing number of national and international lawmaking defines the access to the information in a Plain language or Easy-to-Read as a matter of democracy and inclusion (Matausch, Peböck, & Pühretmair, 2014).

Web content might be of various kinds such as text, images, sounds, videos, and animations. Due to the increased globalization, the economy of the world has become more integrated by which information technology plays a major role in it (Alampay, 2006). Therefore, an image can contain the information which might be of global interest, and that indicates the importance of image accessibility on Web. Images again might be of several types: informative images, decorative images, functional images, images of text, complex images, groups of images, and image maps (W3C, 2014b). Accessibility of these images simply means if the intended information given in these images is reachable and understandable to the disabled people such as visually impaired people, and the normal people (Eggert & Abou-Zahra, 2014). According to the Web Content Accessibility Guidelines (WCAG 2.0), images can be made accessible through image description which is also known as alternative text (ALT text). The importance of accessible image description is not only for the disabled people like visually impaired people but also for those users who don't have high internet speed and have disabled the image display options. In addition, it is beneficial for Web content owners also because search engines such as Google and Bing mainly use image description in their search algorithms. Hence, it is important for a wide range of people to have accessible image descriptions to the images posted on Web.

1.1 The Statement of the Problem

Though there are standard guidelines such as WCAG 2.0 to make Web images accessible to everyone, still there is a lack of its use in the real-world scenario. For example, Francis, Al-Jumeily, and Lund (2013) have analyzed e-commerce site (amazon.co.uk) to find out the usability issues for visually impaired people & found the lack of proper implementation of alternative text in case of non-text elements. Likewise, another study conducted by Bavani, Azizah, and Yatim (2010) has revealed Web experience of visually impaired people in Malaysia and came up with the result reflecting inappropriate use and lack of alternative text for non-text contents on Web. On the other hand, the accessibility evaluation review conducted by Gonçalves, Martins, and Branco (2014) for Portuguese enterprise Web sites mirrored the alternative text missing error for non-text elements as one of the top five accessibility errors while counting the number of existing errors. Hence, the literature shows that there is a significant level of lacking in the effective use of descriptive summary for non-text Web contents. This study did not find any specific research to explore the reason behind this but Bigham, Kaminsky, Ladner, Danielsson, and Hempton (2006) have stated that the negligence of Web authors and the complexity of construction and verification of alternative text causes the increment of inaccessibility of non-text Web content. This motivated the study to explore the existing solutions and propose a more effective method in order to encourage Web users to author quality image descriptions to make them accessible.

There are some studies being done that focus on generating and managing equivalent alternative descriptions for non-text contents. Open source image description tool, POET (BeneficentTechnology, 2017) is one of the existing solutions and is focused on generating descriptions for images that are available on books. On the other hand, there is a research conducted by Demir et al. (2010), which produces description summary for simple bar chart automatically. Furthermore, Doush, Pontelli, Son, Simon, and Ma (2010) have introduced a system to present two dimensional chart through multiple modalities—haptic, sound and visual. Several other researchers (e.g., Xiao, Xu, & Lu, 2010; Yu, McAllister, Strain, Kuber, & Murphy, 2005; Rodr et al., 2015) have conducted studies regarding assessment, management, and development of the

alternative summary text. However, none of these existing solutions have generated accessible image descriptions for images including complex graph and map images.

1.2 Study aims

Authoring Tool Accessibility Guidelines (ATAG) (W3C, 2005) suggests that the tool itself should be accessible for disabled developers, at the same time it should facilitate developers to produce accessible content. Hence, this study investigated on how sample cues can simplify writing image descriptions even for general people who may or may not have the previous experience on describing images in order to make them accessible to visually impaired and blind people, and at the same time improve the quality of the description for a better accessibility. To achieve this goal, this study has come up with the following research questions.

1.3 Research Questions

1. *Can we simplify and improve quality (in compliance with NCAM guidelines) of image descriptions for accessibility by providing sample cues?*
2. *How are the difficulty level and time it takes for writing image description affected by a sample cue and different sample cues?*

1.4 Thesis Organization

This thesis consists of six chapters including this Introduction chapter. Chapter 2 presents the background that includes relevant concepts and related research. Chapter 3 presents the cue-based method for describing images and methods for data collection and analysis used in this thesis. Chapter 4 incorporates experimental setup and steps together with the results. Chapter 5 discusses the results in relation to previous research and explores the possible reasons for the results found in this study, reflects the strength and weaknesses together with the usefulness of the study. Chapter 6 comes up with concluding remarks and shows the possible way further to the future studies. Reference list and appendices can be found after the final chapter.

2. Background

This chapter examines related research and concepts on image accessibility. First, it provides a broad picture of universal design concepts in relation to image accessibility in *Section 2.1* which incorporates the concept of Web accessibility in relation to the universal design of ICT, image accessibility on Web, and people's limited abilities in relation to image accessibility including both physical and situational disabilities. *Section 2.2* investigates several social aspects of image accessibility. Technical aspects are essential to be discussed to solve the image accessibility problem on Web and that is discussed in *Section 2.3*. After investigating different aspects of image accessibility, *Section 2.4* explores several guidelines for accessibility including NCAM guidelines used in this study for image description evaluation. *Section 2.5* explains the importance and implementation of image description in detail. Existing image accessibility solutions are incorporated in the last section i.e. *Section 2.6* which reflects both human-based and computer algorithm-based solutions, explores why they are not enough, and propose a cue-based solution for describing the image in brief.

2.1 The Concept of Universal Design

The term universal design came from the North Carolina University during the late nineties, and formed a movement of students and researchers, in order to make interior and exterior design easier to use for disabled individuals (Story, 2001). Promoters for this design explain universal design as "The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaption or specialized design." The same promoters are also accountable for creating the seven principles of universal design:

- equitable use,
- flexibility in use,
- simple and intuitive use,
- perceptible information,
- tolerance for error,

- the low physical effort,
- size and space for approach and use (Story, 2001).

Universal design principles were created to articulate the full range of limitations for achieving universal design for all kinds of designs (Story, 2001). In addition, they elucidated how the concept of universal design might affect specific designs under development. Furthermore, they suggested how usability of those designs could be maximized. The resulting principles, mentioned above, have their own guidelines to achieve these principles (Story, 2001). For example, to make a user interface simple and intuitive, it needs to be: without any un-necessary complexity; consistent with user expectations and intuitions; consistently arranged to give information with its importance; and providing effective feedback after each of the task completion (Story, 2001).

The overall goal of universal design is to design for all people so that there will be no necessity to retrofit or make the design accessible at a later stage. Universal design is not only about an adaptation but also has the goal of including all people, to the greatest extent, in the design process to begin with. The goal of universal design is also argued to be more cost-efficient and less time consuming than the accessible goal that designs for specialized users, e.g. deaf, visually impaired etc., because it does not cause any usability barriers to begin with and will therefore avoid post-design costs and configuration (Lazar, Goldstein, & Taylor, 2015; Maisel, 2010; Trewin, Cragun, Swart, Brezin, & Richards, 2010; Wentz, Jaeger, & Lazar, 2011).

2.1.1 Web Accessibility. In the context of Web, typical variation between universal design and Web accessibility is that Web accessibility does make sure if websites are usable for users with disabilities with or without assistive technology. On the other hand, the universal design ensures that the possible widest user base—including age, language, culture, disabilities etc.—can use the solution.

Maisel (2010) argues that users can be unable to perform or participate in activities caused by environmental barriers, or on the other hand be enabled by environmental facilitators. The universal design avoids placing environmental barriers in the first place and ensures the usable environment for everyone in the society (Maisel, 2010).

Mpofu and Oakland (2009) further suggest that the designs which are done for specific users, for example, accessible design, are not necessarily universal. However, while all accessible design is not universal, all universal design is accessible.

Erkilic (2011) says that universal design "...is originated and developed within the discourse on disability...". In addition, legislation that supports accessibility and universal design, both in terms of physical space and ICTs, often emphasizes on people with disabilities (tilgjengelighetsloven, 2008; UN, 2006). Through "disability divide", Dobransky and Hargittai (2006) and Solomon (2000) define a gap between a user's ability and the required ability of technology. Furthermore, Organization (2001) defines disability using the medical model, that is a characterization of a person directly caused by trauma, disease or other health condition. Though there are several types of disabilities, the scope of this study is image accessibility on Web, so the following section explains the types of disabilities that make relation to accessing an image on Web.

2.1.2 Image Accessibility. According to the World Health Organization's (WHO's) estimation, 285 million of people affected by visual deficiencies, among which 39 million are totally blind (Awada, Issa, Ghannam, Tekli, & Chbeir, 2012). Because of the loss of certain color information, regions or objects in several images cannot be recognized by these viewers and this may degrade their perception and understanding of the images (Wang, Sheng, Liu, & Hua, 2010). The situational and physical disabilities in relation to image accessibility are explained in the further sub sections.

2.1.2.1 Types of Disabilities in Relation to Image Accessibility. Impairments in the senses are covered by sensory impairments, for example, sight, hearing, smelling, tasting, the sensation of touch and balance etc. Blindness, color-blindness, deafness, and contrast sensitivity are among the impairments which make difficult, or impossible, to perceive visual (image, text, video etc.) or auditory feedback (Lazar et al., 2015). Accessibility measures that can be implemented to ICTs for these kinds of impairments encompass alternative methods of guides or feedback, for example: in case of image accessibility, the alternative method could be the tactile images and text description to the image that can be read by assistive technology (Lazar et al., 2015).

Impairments in skills such as concentrating, thinking, reading, writing, and reasoning lie under cognitive impairments. Dementia, Autism, Dyslexia, and Attention Deficit Hyperactivity Disorder (ADHD) are some examples of such impairments. In the context of ICT and specially in image accessibility, cognitive impairments make difficult for users to read or comprehend image description. It is even more difficult in case of description for complex images such as a graph, map etc. On the other hand, it is also harder to write descriptions to the complex images by the people with cognitive impairments. Accessibility measures that can be implemented to ICTs for these kinds of impairments contain definitions or explanations for: unusual words and jargon, easy-to-read textual information, correct use of graphics and graphical changes, customizable and intuitive user interfaces and other assistive technologies such as screen reader software etc. (Karger & Lazar, 2014; Lazar et al., 2015).

2.1.2.2 UD and Image Accessibility in Non-Disabled Scenario. For many disabled users, accessibility to a technology is a requirement. However, non-disabled users can also be facilitated from universal design and increased accessibility. There are considerable examples which show how universal design has benefitted non-disabled users. Some commonly known inventions that emerged because of the universal design movement are the automatic door openers, the “dropped curb” etc. The concept of automatic door openers was initially for wheelchair users or walkers to make ease entrance. Later on, other people such as people carrying grocery bags or babies also benefitted because of it (Burgstahler, 2004). The dropped curb, designed by architect Selwyn Glodsmith in the 1960s (Warschauer & Newhart, 2016), was an invention to enhance wheelchairs to enter the sidewalks from the street by making an angled ramp on selected places of the curb. Afterwards, this invention became useful for baby trolleys, bikers, and skateboarders etc. There are many other examples of inventions related to UD which can be read in (Fuglerud & Sloan, 2013).

Likewise, in case of image accessibility on Web, the accessible image is not only useful for blind or visually impaired people but also for those people who do not have sufficient internet speed to download the actual image, and the people or company who want their images findable on Web through search engines like Google and Yahoo.

2.1.2.3 Situational Disabilities and Image Accessibility. Lin and Seepersad (2007) suggest situational disabilities as the conditions where ordinary users function in extraordinary environments. These situations cause a state of temporary impairments due to environmental factors to the users who may or may not have an existing cognitive, motor, and sensory disability. For example, a noisy environment may cause a significant barrier in the hearing, difficult to concentrate under stress, or even a user without a motor disability may not be able to touch the screen while holding onto handrails in a train or bus. Lazar et al. (2015) provide an example of how closed captioning of TV broadcast is useful for both people with hearing impairments and persons who are in cafes or gyms where external noise might dominate the original audio from the TV broadcast.

In the same way, the people without any disabilities might not be able to access Web images in some situations, for example, people might not have enough mobile data to turn on the image display functionality, and places having low internet coverage might not support to display an image on the device. Likewise, in the glare of sunlight, one can unable to recognize images shown on a device screen.

2.2 Social Aspects of Image Accessibility

This section explains how an accessible image performs important roles in order for achieving equal access in the information society. Furthermore, it explores diverse Web content authors in order to clarify how non-textual contents are produced on Web.

2.2.1 Information Society. Modernization theory suggests that the modernization of societies happens in a series of phases and stages having a base for production (Alampay, 2006). For example, the industrial society was characterized by industry, and how machines, factories, and corporate made up the society and everything within. Likewise, the information society have had, and still has, information (technology) as a basis.

2.2.2 Digital Divide. The digital divide is known as the difference in access to ICTs on a global scale. It is defined as “situations in which there is a marked gap in access to or use of ICT devices” (Campbell et al., 2001). Particularly, investigators have shown evidence that new ICT solutions do not consider this digital divide and the level of ICT access between men and women (Richardson, Ramirez, & Haq, 2000), poor and rich

(Gomez, Hunt, & Lamoureux, 1999), rural and urban (Campbell et al., 2001), and people with different education levels (Madhusudan, 2002; O'Farrell, 2001).

Researchers have offered the idea that ICTs are tools for accessing information, communication opportunities, and knowledge (Kirkman, 2000). Taking this idea further, people's possibility to consume information and social participation will be deprived by unequal access to ICT, and therefore people will be unable to involve in a society's development (Helbig, Gil-García, & Ferro, 2009).

Nevertheless, researchers claim that the idea of the digital divide cannot be justifiable and are non-existing, because those who need ICTs have them, and those who do not need ICTs do not have them (Warschauer, 2004). In the meantime, other researchers give strong evidence that access to ICTs can make a difference to people who have been deprived of it (Ching, 2004; Goldstein & O'connor, 2000). Furthermore, three independent surveys, investigating the UK population, reported that "those who are most deprived socially are also least likely to have access to digital resources such as online services" (Helsper, 2008).

Thus, the gap that occurs because of the digital divide may decrease, or even disappear, with universal service universal access. Verhoest and Cammaerts (2002) explain these terms as based on accessibility, affordability, and quality of service. Universal service emphasizes the availability in all communities. This may make clear about why developing countries struggle for universal access, because of the lack of market and resources, while developed countries strive for universal service (Alampay, 2006).

2.2.3 Diverse Content Authors on Web. Due to the proliferation of internet technology, smart phones, and social networking sites such as Facebook, Twitter etc., many people— including Web developers and content producers working under several companies— have become potential Web authors who post images, videos, and text content on Web.

According to Facebook (2017), it has 2.01 billion monthly active users as of 30th of June 2017, who are posting contents on the Web through the Facebook user interface. Likewise, 330 million of people posting their tweets per month during the 3rd quarter of 2017 (Statista, 2017a).

The users of WordPress, which is a website and blog engine, produces 81.8 million new posts and 44.9 million new comments each month (WordPress, 2017). Likewise, Tumblr (2017), which is also a blog engine, has 374 million of blogs, filled with multimedia. In addition, the number of active users per day for Instagram, a free mobile application to share photos and videos privately or publicly, has reached 500 million (statista, 2017b).

These are statistics only for the popular social networking sites and blog engines. There is a long list of many other Web applications and Websites through which people are communicating or expressing their ideas have not been mentioned here. Hence, it gives us the clear impression of how people are being involved in the process of creating multimedia contents on the Web. In the next section, this document will explore existing Web authoring tools and their user interfaces through which Web workers and other users post contents on the Web.

2.3 Technical Aspects of Image Accessibility

Here, the scope of technical aspects of image accessibility is limited to the user interface of several authoring tools and popular Websites which are used on Web in order to post contents including images. This section explores how the current Web interfaces and authoring tools such as Content Management Systems (CMS) are not supporting normal users and Web workers writing an accessible image description.

2.3.1 Web Authoring Tools. Authoring tools are defined as software and services that help authors—Web developers, designers, writers, etc.—to produce Web content such as static Web pages, dynamic Web applications, etc. (W3C, 2005). According to W3C (2005), some examples of authoring tools are:

- web page authoring tools, such as what-you-see-is-what- you -get (WYSIWYG) HTML editors.
- websites generating software—CMS, courseware tools, content aggregators etc.
- software that adapts to Web content technologies, for example, word processors and other office document applications with “Save as HTML”.
- authoring tools for multimedia.
- websites that support users add content, for example, photo sharing sites, blogs, social networking sites, and online forum.

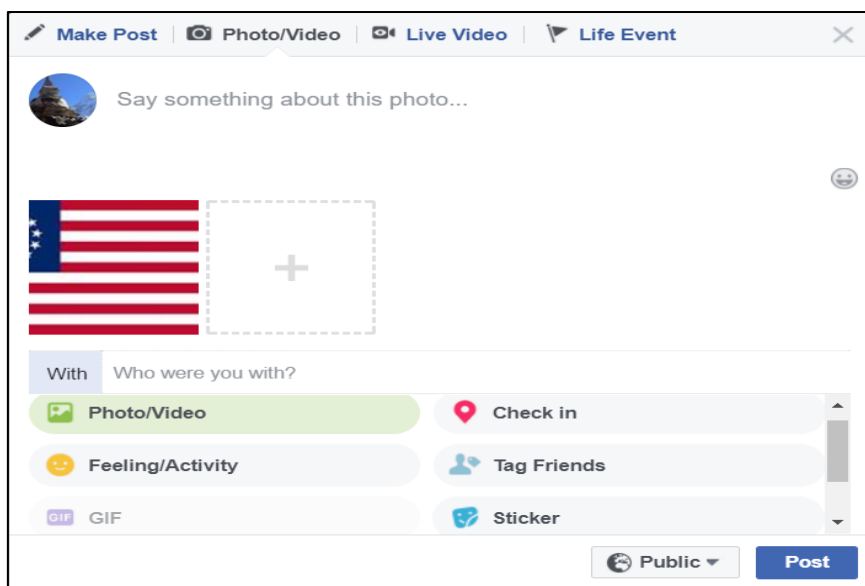


Figure 2.1. A graphical user interface to upload a post with an image on Facebook.

Figure 2.1 shows the Graphical User Interface (GUI) for uploading a post on Facebook. Though it has provided multiple options for users such as check in location, tag friends, feeling/activity, and sticker, however, it does not provide any option for image description, so the user cannot write about the posted image. Instead, to fix the alternative text problem, Facebook has launched Automatic Alt-Text (AAT) in 20 languages. Each description starts with “image may contain” followed by the concept tags (Wu, Wieland, Farivar, & Schiller, 2017).

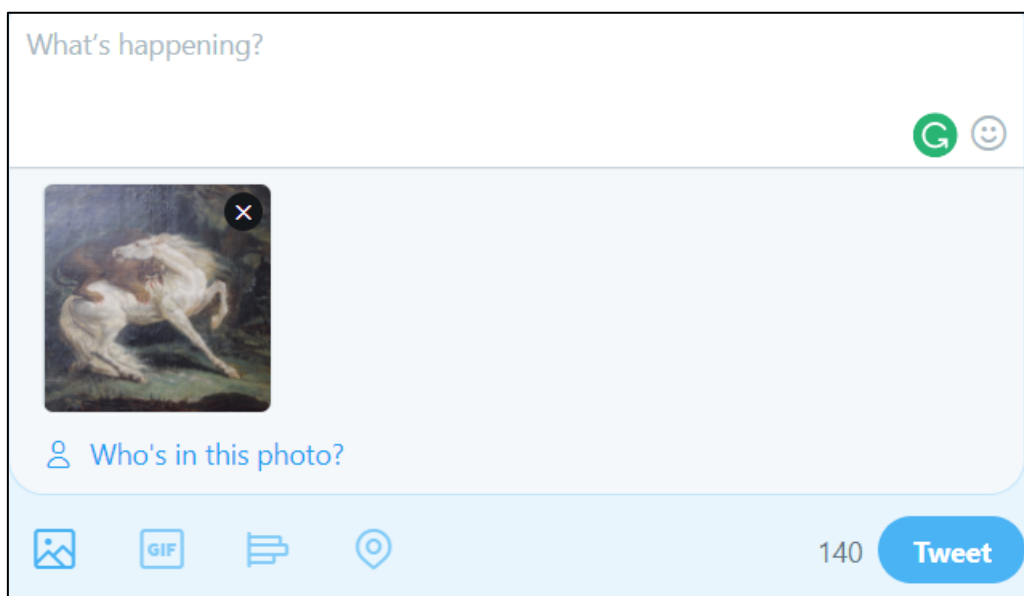


Figure 2.2. A graphical user interface to upload a post with an image on Twitter.

Figure 2.2 shows the GUI for posting a tweet on Twitter. It provides a user option for writing text (up to 140 characters), uploading image together with the option tag people who are in the pictures. But it does not provide any space for describing the image as an alt text. Some users may choose to use the 140 characters of their tweet to explain an embedded image; and which is quite rare, with only 11% of multimedia tweets having text that can be used as image description (Morris et al., 2016). However, one grassroots effort to retrofit alt text into tweets is the Alt Text Bot. It is based on Cloudsight API. When a Twitter user forwards a tweet with an embedded image to the Alt Text Bot's account, the API service tweets back a short caption (Morris et al., 2016).

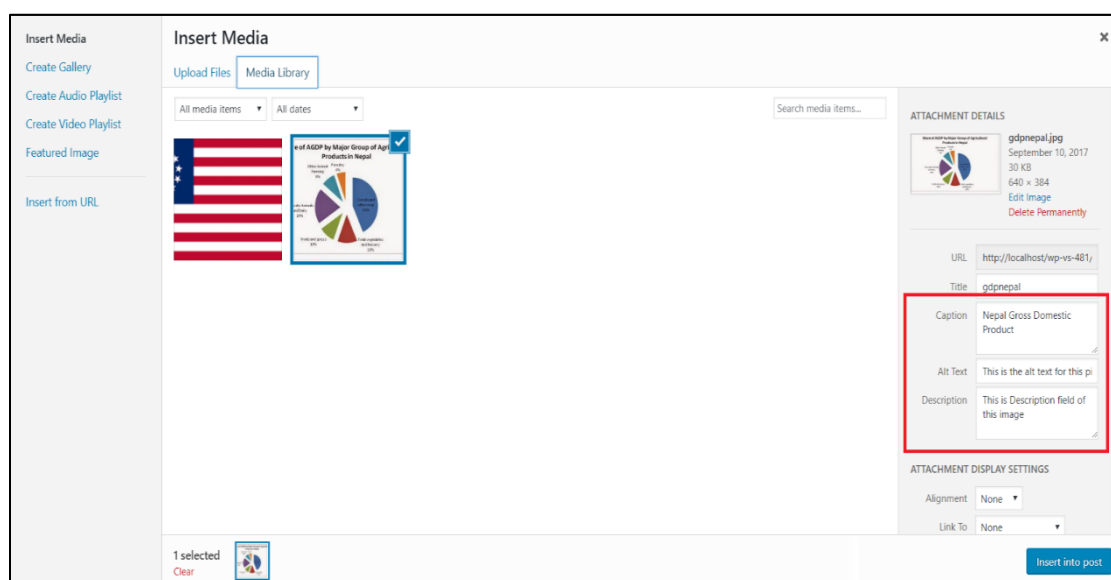


Figure 2.3. A graphical user interface for inserting image to a page as a content in WordPress.

Figure 2.3 shows the GUI for inserting image to the Web page from media library in WordPress. Here, the Web content author is provided a simple form on the right side of the page to fill up the attachment details such as URL, title, caption, alt text, and description.

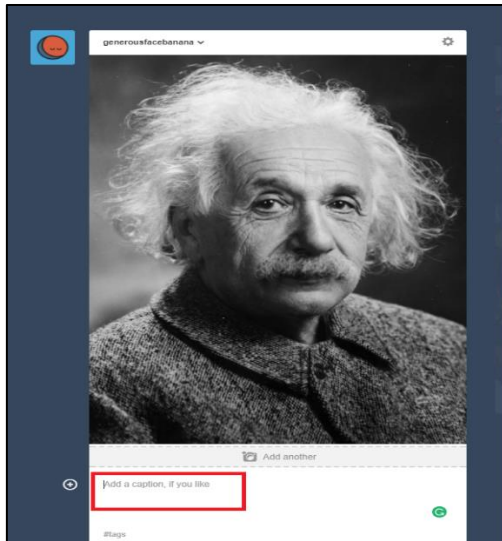


Figure 2.4. A GUI for uploading an image on Tumblr blog.

Figure 2.4 shows the GUI for uploading an image in Tumblr blog (Tumblr, 2017). It provides a user option for adding a caption if they like.

2.3.2 Content Management Systems and Accessibility. There are several reasons for businesses not achieving the universal design in practice. (López, Pascual, Menduiña, & Granollers, 2012) suggest that a content management system can be a bridge between Web content and creator, regardless of creators' background whether they are from computer technical background. In other words, the purpose of authoring tools was to remove the technical details, code, markup to bloggers, and news organizations together with other content creators (Harper & Yesilada, 2008). One might argue that the authoring tools are immensely important in case of content creation because of this fact. Due to this reason, researchers have explored the effect accessible authoring tools have on Web content, and to what limit its ability to facilitate the content creator, encourages a universally designed product and compliance with industry guidelines.

Lazar et al. (2015) say the tool for creating content such as a CMS or learning management tool (LMS) should be accessible and able to create accessible content for the users.

As a disappointing starting point, Freire, Russo, and Fortes (2008) claim that there is a very low number of CMS developers who take into consideration the, industry standard, Authoring Tool Accessibility Guidelines (ATAG), see *Section 2.4.2*, or other

similar guidelines during developing authoring tools. In consequence, the authoring tools lack accessibility features and also discourages or at worst case removes, the ability to create accessible content in the authoring tool.

Bittar, do Amaral, Faria, and de Mattos Fortes (2012) compared Dreamweaver; Eclipse —Helios; Netbeans 7.1; Microsoft Expression Web and NVU 1.0 through guideline compliance test. In this comparison, they used scale having different numbers such as 0 for “does not comply”, 0.5 for “partially complies” and 1 for “complies”. As a result, they found that none of the tools collectively met with WCAG 2.0 criteria, see *Section 2.4.1*, however, some of them complied in varying degree, individually. Furthermore, they came up with the evidence that might signify that authoring tools which are paid or specialized for Web development complied the selected criteria at a higher degree compared to the free or open-source ones. In addition, none of the authoring tools created accessible tools, and very low number made the user aware of the use of e.g. heading titles, link titles, and structure.

Likewise, Pascual, Ribera, and Granollers (2012) conducted an evaluation for two other popular CMS-systems, Blogger and WordPress, to figure out if they comply with ATAG 1.0 and WCAG 1.0. They got the result for ATAG 1.0 that shows Blogger and WordPress failed on 71.43% and 53.57% respectively. Similarly, for WCAG 1.0, they also found the results. In addition, both content management systems, with their default settings, did not comply with a single one of the ATAG priority one requirements. Particularly, both systems showed two main problems: the lack of UD promotion offered by the authoring tool and the creation of inaccessible content. However, this study did not check system conformance with the current version of WCAG 2.0. But, the research still shows that ATAG conformance relates to WCAG conformance—as intended by the guideline authors.

2.3.3 Assistive Technology. The term ‘Assistive Technology’ (AT) mentions to the set of scientific accomplishments—products, environmental modifications, services, and process—beneficial to overcome restrictions and/or expand function for an individual (Cook and Polgar, 2014). In other words, AT helps persons with disabilities or special needs to deal with their daily activities in order to achieve a better quality of life (Lancioni et al., 2012). Among the various functional groups of AT related to users’

needs, the *Sensory Functions* deal with the sensory impairments having reduced ability (or lack of ability) regarding vision, touch, and hearing senses (Leo, Medioni, Trivedi, Kanade, & Farinella, 2017). Because of the computing capabilities, it has been possible to build powerful and diverse applications in this context. More specifically, text detection and recognition for text-to-speech AT can help people with severe vision impairment (Leo et al., 2017). However, there are some core issues for each user group that does not truly change even if the technology does. For example, blind people need to be able to access equivalent description that gives an idea of what a particular image is about (Connor, 2012).

2.4 Guidelines for Accessibility

This section explains several guidelines related to image accessibility. It focuses on the NCAM guidelines this study chose to evaluate the image descriptions collected from the participants.

2.4.1 Web Content Accessibility Guidelines (WCAG). Web Content Accessibility Guidelines (WCAG) 2.0 includes a broad range of recommendations for making Web content more accessible. These guidelines support to make content accessible to a varied range of people with disabilities, including blindness and low vision; deafness and hearing loss; learning disabilities; cognitive limitations; limited movement; speech disabilities; photosensitivity and combinations of these. Regarding non-text content — image, videos, animations etc.— accessibility, the guidelines say that all non-text content that is presented to the user has a text alternative that serves the equivalent purpose except for the situations such as if the non-text content is ‘controls’ or ‘input’ etc. There are other situations which can be read on (Ben Caldwell, Cooper, Reid, & Vanderheiden, 2008). Although there are specific guidelines for Web content accessibility, Kelly et. al (2013) claims that Web accessibility practices and policies need to be flexible and should consider context as an important factor.

WCAG 2.0 describes the way of making Web accessible to everyone. Specifically, regarding image accessibility, it gives emphasis on alt/long text which is equivalent to the content of non-text content, for example, graph images (Ben Caldwell et al., 2008).

According to Consortium (2008), the Web Content Accessibility Guidelines is a set of guidelines that “covers a wide range of recommendations for making Web content more

accessible. Following these guidelines will make content accessible to a wider range of people with disabilities, including blindness and low vision, deafness and hearing loss, learning disabilities, cognitive limitations, limited movement, speech disabilities, photosensitivity and combinations of these.”

However, Aizpurua, Arrue, Harper, and Vigo (2014) showed that the WCAG 2.0 guidelines, particularly with the use of automatic tests, did not cover all the accessibility issues given by users of a web page. Initially, the researchers differentiated the areas of a Web page where the user experienced problematic situations. The differentiation was run through gathering evidence of a participant’s coping behavior while performing a task. Afterwards, they made algorithms from the user’s coping tactics through translating the behavior into machine-readable code, by which the behavior could be simulated. Finally, these algorithms were deployed to the Web page. In addition, they also advised expanding the algorithms to notify experts, web page owners, and the researchers. The researchers recommended that this method would be far better than guidelines, because of its user and product specific nature.

Power, Freire, Petrie, and Swallow (2012) conducted an empirical study of blind users’ problems on the Web. The study presented that only 50.4% of the problems were addressed by the WCAG 2.0 guidelines. Furthermore, the study reflected that the web pages that implemented techniques to achieve the WCAG 2.0 Level A did not actually give any solution to the experienced problem. Because of this, the researchers recommended a design principle approach over a problem-based approach in order to get a higher degree of accessibility.

Cooper, Sloan, Kelly, and Lewthwaite (2012) mentioned that W3C Web Accessibility Initiative (WAI) and its guidelines for authoring tools and websites are more product oriented, and not sufficient for requirements and user goals. They argued that the Web workers with only recent knowledge of technical guidelines and properties are not met with the social aspect of Web accessibility between product and user. Furthermore, they stated that the technical properties and usability metrics are not only the factors that confirm the accessibility of the internet. However, there are other factors also such as economic, social, and political aspects. In addition to this, they mention that staff, end-users, and processes should play a bigger role than before during the development of

Web sites to ensure a higher degree of accessibility and this concerns internal operations in the companies or organizations that produce accessible Web solutions. The authors presented the evidence that reflected the importance of standards that focus on the best process and practice to develop a much more accessible end product.

2.4.1.1 Alternatives to WCAG. Cooper et al. (2012) have shown further accessibility initiatives, for example, the IBM Social Accessibility Project and Fix the Web, which supports the users to inform their experiences without the need to mention the technical underlying aspect. However, these systems are still not widespread and scale up to the extent the initiators, in the beginning, wanted them to.

2.4.2 ATAG. Authoring Tool Accessibility Guidelines (ATAG) is a standard, recommended by W3C and developed by the Authoring Tool Accessibility Guidelines Working Group (ATAG WG). It contains the requirements of the authoring tool that need to be fulfilled in order to produce accessible Web contents. It has been updated from ATAG 1.0 to ATAG 2.0 (W3C, 2005).

According to the research, the authoring tools have to comply with the Authoring Tool Accessibility Guidelines (ATAG) 2.0 in order to produce the accessible content using them (Treviranus, 2008). The authoring guidelines cover two aspects of authoring tools and they are: making the tool itself accessible and supporting the creation of the content which is accessible (W3C, 2005).

2.4.3 NCAM Guidelines. These are guidelines which enable image describer to describe several types of images such as map, graph, and general image. They were made by the Carl and Ruth Shapiro Family National Center for Accessible Media at WGBH (NCAM) together with the DIAGRAM Center (Digital Image And Graphic Resources for Accessible Materials) at Benetech (NCAM, 2009).

These guidelines are made on a multi-study project, which performed two rounds of a Web based Delphi survey, taken by more than 30 expert describers and individuals with vision loss, to establish approaches to the description of Science-Technology-Engineering-Mathematics (STEM) images. A follow-up 60 persons end-user study, with participants who had visual impairments, confirmed that the description guidelines produced quality image descriptions, with higher clarity and efficiency (Morash, Siu, Miele, Hasty, & Landau, 2015). The guidelines were initially focused only on STEM

images used in digital learning materials. Later, it incorporated general best practices that apply to all types of images, as well as, an expanded set of image-specific recommendations. The expanded recommendations contain the image types frequently found in the humanities and social sciences such as maps, photographs, and art (NCAM, 2009).

2.5 Describing Image

This section talks about why describing the image is important. It reflects how writing a short text for an image is an artistic work. Finally, it informs the possible reasons people do post images on Web without textual descriptions.

2.5.1 Importance of Image Description. Several countries' official sites — USA.com, Statistica.no, Canada.com etc. — use images and diagrams in a significant level to demonstrate diverse statistics regarding population/races, income, crime rate, education, housing etc. Social service information and statistical data provided on Web are vital for every individual irrespective of his/her abilities to interact with computing since they live in the same society.

The graph images provide high level information regarding data trends and statistical variations, so it is not easy to read and understand the intended messages within it. Specially, people having cognitive disability might experience a greater level of difficulties to interpret chart (W3C, 2014a). It is therefore beneficial to provide alternative description not only for low to no vision people but also cognitively disabled people and even normal people as well. Hence, equivalent alternative description increases accessibility as well as usability of images on Web.

There have basically three different interfaces been evolved to make the image accessible on Web: haptic, audible, and text description. The 'text description' equivalent to graph image explains only the most important information precisely and in a consistent order which supports user to anticipate further information that is going to be provided. According to NCAM guidelines (NCAM, 2009), STEM images can be best described by linear and narrative description considering brevity, drill-down organization, clarity and emphasis on data.

It is important to provide image description because not all user may have the ability to "see" the images, for example, blind people, and screen readers can not interpret

images. The screen readers must rely on text to read out loud the given information on the page to people who are blind as well as others who use them. The non-disabled users on slow dial-up may decide not to download images due to slow transfer times. Furthermore, when content is necessary to be enlarged to be seen, the text often scales better than images, which can become pixelated and/or can easily take over the entire screen, making users have to scroll the image vertically and/or horizontally (Stanford-University, 2015).

In addition, image description as an alt text is even more important for Website owners because Google cannot see images but can read equivalent alt text which increases search engine optimization (SEO) index—measures popularity— of their Websites. Stanford-University (2015) further claims that text allows information to be presented in other formats, for example, tactilely or in ways, we don't even know about yet.

2.5.2 Art of Writing Alternative text. According to Slatin (2001), an ALT text is a phrase or sentence “attached” to an image or other elements so that people who use assistive technologies such as screen readers and talking browsers, in addition, people who prefer to browse the Web turning off images, can identify the element. Furthermore, Slatin (2000) explains ALT text functions as (a) it supports to identify briefly the non-textual element to which it is attached, and (b) provides access to the functionality represented by that element.

In addition, because the participants who need ALT text are usually in no position to judge equivalence—since they don't face the original in the first place— therefore, the ALT text should be accurate, descriptive, and concise (Slatin, 2001). This study did not find any explanation regarding the length of ALT text in WCAG 2.0, however, the theoretical upper limit is over 65,000 characters (Slatin, 2001). On the other hand, assistive technologies may implement *de facto* limits, for example, an obscure setting in the JAWS screen reader causes problems if ALT text exceeds 150 characters. However, users can change this setting (Slatin, 2001).

Thus, producing effective ALT text is an exercise in extreme economy. The challenge is to encompass as much information as possible in the minimum characters without compromising intelligibility (Slatin, 2001).

2.5.3 Image Description Implementation. The negligence of web authors in providing alternative text is one of the causes of web inaccessibility, in addition, the proper selection of alternative text is considered by many to be more of an art than a science, which increases the difficulty of construction and verification of alternative text (Bigham et al., 2006). Furthermore, it takes a long time to read image description guidelines for Web developers and they don't want to spend their time on it. Many of Web development tools don't support features to write image description since they work on drag and drop basis. Likewise, Takagi, Harada, Sato, and Asakawa (2013) explain writing image description as a difficult process, so developers do not be motivated to write it.

Morash et al. (2015) say explicit instructions on image description guidelines is not sufficient to produce quality image descriptions when using novice Web workers. Instead, it is better to provide information about images, then generate descriptions from responses using templates. Hence, it is better to provide real time guidance rather than traditional guidelines in order to write image description effectively. The further section explores existing human-powered and algorithm-based image accessibility solutions in detail.

2.6 Related Works on Image Accessibility

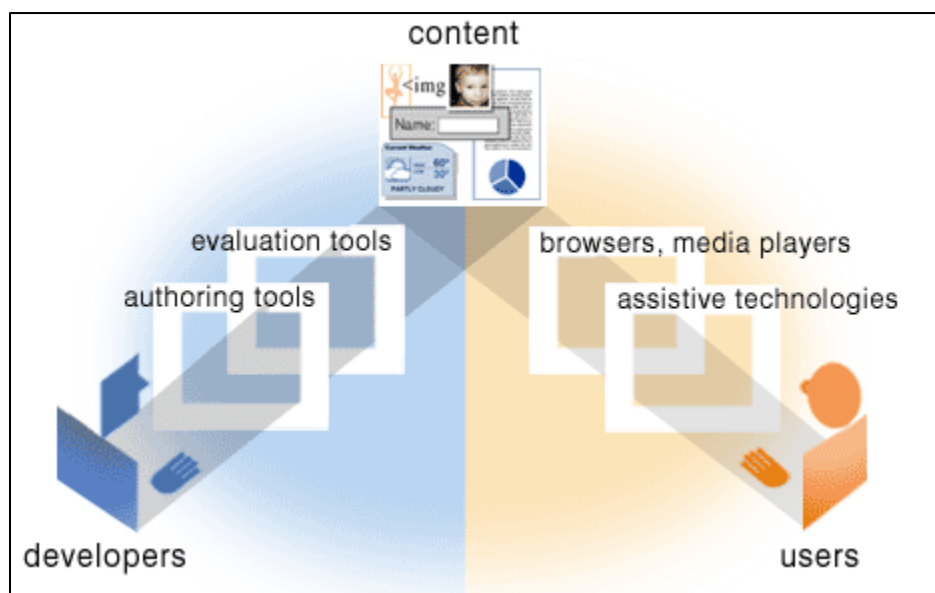


Figure 2.5. Related components of Web accessibility (W3C, 2005).

Figure 2.5 shows the related components of Web accessibility. Web developers or Web authors use authoring tools to produce Web contents. Furthermore, they use evaluation tools, which perform evaluation based on WCAG standards, to make sure if the produced materials are accessible. On the other hand, users browse the content through user agents such as browsers and media players. Assistive technologies, for example, screen readers help to increase users' ability to access the content. The next section of this study explains the related work regarding authoring image descriptions using human powered authoring tools and automatically generated image descriptions. It presents the arguments to support why they are not enough in order to not to have inaccessible images on Web.

2.6.1 Human Powered Authoring Tools for Image Description. A considerable number of researches have been conducted in order to make images accessible to visually impaired such as blind people. Bigham et al. (2010) have contributed a system which helps blind people to get real time feedback for the images they have taken by their talking mobile. The system called VizWiz lets blind people take a picture, ask the question, and receive answers from distance workers almost in a real time.

TapTapSee (2014) is a mobile camera application developed particularly for the blind and visually impaired iOS users and powered by the CloudSight.ai image recognition API. Through the device's camera, VoiceOver functions to take pictures of any two or three-dimensional objects, define and analyze within seconds, and speaks the identification audibly to the user. In addition, it includes other useful features such as repetition of the last image's identification, ability to upload images from the camera roll, share identification through Twitter, Facebook, text or email etc.

Splendiani and Ribera (2014) suggested a method of using decision tree that may reduce ambiguity and enhance the relevance of alternative texts to non-textual elements. The method includes a little modification in the existing task of writing image, table, graph, video, and figure captions. With following a decision tree in a "checklist-like" manner, the content authors can make the most out of the caption. The authors argue that the cognitive load needed to analyze the decision tree is to be less than having to consider previous knowledge due to the visual representation and can ultimately save time in the decision-making process while still developing relevant and

rich figure and image captions. Furthermore, the researchers mention that this method does not interfere with existing workflow and the method is also suggested to be applicable in a Web context.

Zhong, Lasecki, Brady, and Bigam (2015) have introduced a system called RegionSpeak which provides combine visual information across multiple images through image stitching. Furthermore, the system helps to collect labels quickly from the crowd for all relevant objects included within the resulting large visual area in parallel, and interactively explore the spatial layout of the objects that were labelled.

There is a study conducted by Morash et al. (2015) which compared two methods— queried image description (QID) and free- response image description (FRID) method based on NCAM guidelines— for novice Web workers to produce image descriptions for graph images. In QID, series of questions were asked of the participants regarding information available in an image and the text description was constructed satisfying the template with the information given by participants. But, the participants were not informed about the guidelines explicitly. On the other hand, in FRID, the participants were provided images to be explained, together with the NCAM description guidelines. This study suggested QID as the best method, among two methods, to generate consistent and effective image descriptions in case of the graph.

However, the existing human powered systems are constrained by scalability, latency, cost, and privacy concerns (Wu et al., 2017). In addition, none of them supports a Web user who may or may not have previous experience to describe an image that is going to be posted on Web. The next section examines the algorithm-based image accessibility tools to figure out if the existing solutions are enough to provide an accessible text description to an image on Web.

2.6.2 Computer Algorithm Based Image Accessibility tool. Cundiff (2015) has developed a browser extension that adds descriptions to images on the Web for blind people. After getting the user clicks on an image, the extension sends the image URL to the Cloudsight API. During the request is in the process, the image is indicated as “busy” so screen reader users are informed that description is on the way. When the API responds, the extension attaches the resulting description to the image that was clicked, and the description is read by the screen reader.

Farhadi et al. (2010) have demonstrated a system that can compute a score linking an image to a sentence. The computed score, which is obtained by comparing an estimate of meaning obtained from the image to one obtained from the sentence, is used to attach a descriptive sentence to a given image, or to get images that illustrate a given sentence.

Ramnath et al. (2014) have introduced a system that supports a smartphone user generate a caption for their photos. It is based on cloud service where several modules are applied to recognize a variety of entities and relations. The combined outcomes of the different modules result in a large set of candidate captions which are provided to the phone.

Using Google (2017) people can do a Google search by taking a picture with their device through the Google Goggles app. If Google identifies the image in the photo, the app will tell them what's in the photo.

Jayant, Ji, White, and Bigham (2011) have presented an application called EasySnap which provides audio feedback to support blind people to take better pictures of objects and people.

BeMYEyes (2017) on the other hand is an app that connects blind and visually impaired people with sighted helpers from around the world via video connection. In this app, blind people request assistance to overcome the challenges such as navigating new surroundings etc. and the volunteer helpers receive a notification for help. Through live video connection, the volunteers help the blind people by providing answers to the questions they ask.

Brady, Zhong, Morris, and Bigham (2013) have explored the potential of blind users asking visual questions in social networks. In addition, they also have explored whether blind users find social networking sites (SNSs) suitable for Q&A. To do this task, they used a log analysis of questions asked by using VizWiz (Bigham et al., 2010). Their findings suggested that blind people have a large presence on social networking site, however, do not take them as an appropriate venue for asking questions because of high perceived social costs.

Fang et al. (2015) claimed their automatically generating image descriptions approach as a novel approach which includes visual detectors, language models, and

multimodal similarity models learnt directly from a dataset of image captions. They used multiple instances of learning to train visual detectors for words that usually occur in captions, including several parts of speech, for example, nouns, verbs, and adjectives.

Karpathy & Fei-Fei (2015) presented a model for natural language descriptions generation of images including their regions. Their model was based on Convolutional Neural Networks over sentences together with a structured objective that aligns the two modalities through a multimodal embedding. Von, Ahn, and Dabbish (2004) demonstrated how a game can be used to create labels to the images on Web.

The study conducted by Ferres et al. (2007) has proposed a system called 'iGraph-Lite' to enhance the accessibility of graph information by producing visual description automatically. In this study, they did not use image recognition algorithm to extract the information but rather have used application specific plug-in—written for MS Excel. Furthermore, Doush et al. (2010) have also suggested a similar system which is again based on Open Office XML and Microsoft Excel but have presented the graphical information through multiple modalities—aural cues, speech commentaries, and 3-dimensional haptic feedback. Though these studies address the presentation of graphical information to visually impaired people as an assistive technology, they cannot give any idea about the graph information which is in image form and available on Web pages.

However, Demir et al. (2010) have presented a system called *Interactive SIGHT*, which supports automatic information extraction from simple bar chart given in image form. Based on the extracted information, it generates summary descriptions. While doing so, it, firstly, provides the brief summary of a bar chart and then generates history aware follow up responses to the user requests for the further information about the chart. Though it has tried to address the automation of summary generation for graph images, it encompasses only a simple bar chart. In fact, several other common graphs— such as line graph, pie chart, etc. — are still available there which are beyond the scope of this solution. In addition, this tool does not support other browsers except Internet Explorer.

Several other studies(e.g., Feng & Lapata, 2013; Tariq & Foroosh, 2017) have been conducted to facilitate the automatic generation of image captions as an alternative

text— on the basis of contextual cues—giving an idea about the available images on the text form.

Thus, there exist the significant number of literature regarding generating image description automatically, however, none of them are robust enough to be useful in practical life (Morris et al., 2016). In addition, the descriptions generated through these solutions less possibly comply with the standard guidelines such as NCAM guidelines in order to be understandable and equivalent to the information shown in the image for diverse user groups. Therefore, the next section includes the proposed cue-based image description writing solution.

2.6.3 Proposed Cue-based Image Description Writing. This study came up with a new cue-based image description writing aimed for simplifying and improving effectiveness for accessibility. Web users are provided different example images having quality text descriptions as a sample cue in order to help them in writing descriptions for their images while posting on the Web. The next chapter will discuss this in detail.

3. Cue-based Image Description Writing and Research Methodology

This chapter explains proposed cue-based image description writing in detail along with research methodology and methods for data collection and data analysis used in this study. *Section 3.1* clarifies the basic idea about sample cues, providing sample cues on the user interface, and managing sample cues on Web. The type of research this study falls in is explained in *Section 3.2* which specifically discusses research methodology containing the reason why the particular methods for data collection, management, and analysis were selected together with the information about participants' recruitment, selection of guidelines, and image description evaluation measures. *Section 3.3* includes the pilot study conducted before the actual experiment.

3.1 Cue-based Image Description Writing

The existing authoring tools do not provide the adequate guidance while writing an image description, see *Section 2.3*. In order to address this gap, this study came up with a solution in which different sample cues are provided to the user while posting an image on the Web. In this section, the basic concept of sample cue and providing it on Web user interface is explained.

3.1.1 Sample cue. It is an example image having one or many possible text descriptions that can be read by users in order to understand how the accessible image descriptions look like. Different types of sample cues selected in this study during the experiment are given in *Section 3.2.7*.

3.1.2 Providing Sample cues on User Interface. The basic idea behind this study is to provide a sample cue as a real time guidance on a user interface in order to make

describing images simpler and enhance the description quality in terms of accessibility.

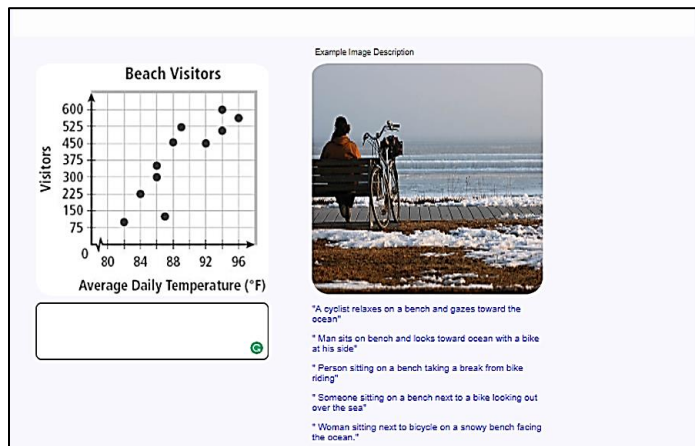


Figure 3.1. A simple user interface having random sample cue while posting a scatter plot graph.

When someone selects an image to post on Web, a sample image, for example, a picture of a cyclist in Figure 3.1, is provided beside the image that is going to be posted, like a scatter plot graph in the above figure. The provided sample gives an idea to describe the image properly in order to enhance the accessibility. In order to provide a sample cue in a real time, it is essential to be manageable.

3.1.3 Feasibility of Providing Similar cues. Since the research in image processing field is growing faster (Remondino, Spera, Nocerino, Menna, & Nex, 2014), modern artificial neural network based methods make it easier to find similar images with high accuracy. Furthermore, the concept of a Web based tool for image annotation (Russell, Torralba, Murphy, & Freeman, 2008) might be useful to produce sample images having accessible descriptions. The produced sample images can be stored in databases or can be managed with cloud service (Dikaiakos, Katsaros, Mehra, Pallis, & Vakali, 2009).

3.2 Research Methodology

Human computer interaction (HCI) is an interdisciplinary setting. Therefore, it is important to examine that why specific methods are suited to particular kinds of work (Hudson & Mankoff, 2014). In this regard, it is possible to distinguish HCI research into two kinds: technical HCI research which is focused on interface building; and behavioral HCI research, which is focused on cognitive foundations. Both of this kinds have their

own expectations in terms of number and background of participants; development of a tool or interface; and outcomes (Hudson & Mankoff, 2014).

This study is intended to investigate how the Web users and authors behavior regarding writing an image description will be affected by providing sample cues while creating the Web content through the authoring tools having a graphical user interface such as CMS and blog. That is why the behavioral HCI research is the appropriate category for this study. According to Lazar, Feng, and Hochheiser (2017), three different types of investigations are possible in behavioral research: descriptive, relational, and experimental investigation. They further have mentioned the descriptive investigation as an act of research where the observations, surveys, and focus groups are used to construct an accurate description of what is happening. Likewise, the relational investigation allows the researcher to identify relations between multiple factors but without the casual relationships between the multiple factors. On the other hand, the experimental investigation can even identify the causal effect between two factors (Lazar et al., 2017).

The main objective of this study is to explore how much does a sample image with proper text description affect in writing image description in terms of the text description standards explained in NCAM guidelines, see *Section 3.2.8*, the time taken, and level of difficulties to write image description. Hence, to come up with some conclusions, it is necessary to determine the casual effect of the time taken to write the description, quality of the description, level of difficulties, and the hints provided as a random cue and similar cue having proper descriptions. Therefore, this study decided to adopt experimental investigation.

3.2.1 Experimental Research. It is originated from behavioral research and covers a broad area of Psychology (Lazar et al., 2017). During the late 19th and early 20th centuries, this approach has highly been accepted in the behavioral Science (Creswell, 2002), and for sure it becomes widespread in the HCI field (Lazar et al., 2017).

Moore and McCabe (1989) suggest that conducting experiment simply means to actively change ‘X’ and observe the response in ‘Y’. In other words, it is about testing an idea to confirm whether it affects an outcome, and with that aspect, the experimental

research design makes possible to identify cause-and-effect relationships (Creswell, 2002).

Furthermore, Creswell (2002) suggests that to conduct a good experimental research, several criteria need to be considered such as random assignment, manipulation of the treatment conditions, outcome measures, control over extraneous variables, threats to validity, and the group comparisons.

According to Lazar et al. (2017) and Key (1997), the experimental research in the field of HCI goes throughout the following processes:

- identify and define the problem,
- generate the research hypothesis,
- specify the experimental design of the study, that represents all the elements, conditions, and relations of the consequences,
- run the pilot study to test the design, the system, and or the study instruments,
- recruit participants,
- conduct the experiment, this is the actual data collection,
- analyze the collected data,
- report the results.

For the experimental investigation, basically, two approaches are available—the online experiment which is also known as Web experiment and the lab experiment. This study chose Web experiment over lab experiment by considering the following advantages:

- web experiments allow the researcher with easy access to a much wider and geographically diverse participant population,
- the natural setting for the participants,
- bringing the experiment to the subject instead of the opposite. It helps participants spare scheduling, transportation, and finding the lab,
- it can be accessed all the time,
- data from Web experiments lend themselves very easily,
- data might be less influenced by interactions between participants' rhythms and levels of the independent variable(s) used,
- participants can choose personally comfortable participation time,

- no limitation to simultaneous use of the materials,
- no scheduling difficulties and overlapping sessions,
- experimenter effect is less,
- it reduces costs (Reips, 2000).

3.2.2 Research Hypothesis. MacLeod-Clark and Hockey (1979) explained a hypothesis as a statement or explanation suggested by observations or knowledge and has not, yet, been proved or disproved which predict the expected outcomes of the research.

In this study, six research hypotheses were generated based on research questions and the relationships found in existing research by reviewing several similar kinds of literature to this study.

1. Hypothesis First

H00: There is no significant effect of random and similar cues on image descriptions in compliance with the overall guidelines.

H01: There is a significant effect of random and similar cues on image descriptions in compliance with the overall guidelines.

2. Hypothesis Second

H0: There is no significant difference in the effect of random and similar cues in compliance with the specific guidelines.

H1: There is a significant difference in the effect of random and similar cues in compliance with the specific guidelines.

3. Hypothesis Third

H0: There is no significant difference in the effect of random and similar cues in compliance with the common guidelines.

H1: There is a significant difference in the effect of random and similar cues in compliance with the common guidelines.

4. Hypothesis Fourth

H0: There is no significant difference in the effect of random and similar cues based on the image types in compliance with the overall guidelines.

H1: There is a significant difference in the effect of random and similar cues based on the image types in compliance with the overall guidelines.

5. Hypothesis Fifth

H0: There is no significant difference in the level of difficulties while writing image description with no cue, random cues, and similar cues.

H1: There is a significant difference in the level of difficulties while writing image description with no cue, random cues, and similar cues.

6. Hypothesis Sixth.

H0: There is no significant difference in the time taken while writing image description with no cue, random cues, and similar cues.

H1: There is a significant difference in the time taken while writing image description with no cue, random cues, and similar cues.

All the hypotheses mentioned above comprise one null hypothesis denoted by H0 and one alternative hypothesis denoted by H1. The null hypothesis informs that there is no relationship between the independent and dependent variables, for example, different sample cues do not influence the quality of image description in terms of compliance with NCAM guidelines. However, the alternative hypothesis is always mutually exclusive with the null hypothesis, that means, if the null hypothesis is true, the alternative hypothesis should be false or vice versa (Lazar et al., 2017).

3.2.3 Research Variables. Variable (s) is the necessary component of the experiment, which has quantity or quality that can vary; participants' characteristics or given study situation which has several values on the study and should be varied or have different levels, is called as a variable (Morgan, Leech, & Barret, 2005). The dependent and independent variables are stated with well define hypothesis (Lazar et al., 2017).

3.2.3.1 Dependent Variables. Level of compliance with NCAM guidelines, time taken, level of difficulties.

3.2.3.2 Independent Variable. Sample cue.

Possible values. No sample cue, random sample cue, and similar sample cue.

Kohavi and Longbotham (2007) say “*a common pitfall in Web experiments is the use of multiple metrics. It's strongly desirable to select a single quantitative measure, or overall evaluation criterion (OEC), to help determine whether a particular treatment is successful or not*”. This is how this study motivated to consider a single quantitative

measure. In addition, Lazar et al. (2017) also support having a minimum number of metrics in experimental research.

3.2.4 Within-Group Design. In this design, each participant to be exposed to multiple experimental conditions (Lazar et al., 2017). Firstly, the participants were exposed to the user interface without any hints, and afterwards, they were exposed to the user interface having random and similar cues simultaneously. This study adopted this design during the experiment by considering the following advantages:

- requires much smaller sample size than between-group design and this is helpful because the qualified participants may be quite difficult to recruit,
- may also help to reduce the cost of the experiments when financial compensation is provided,
- provides an effective isolation of individual differences and the tests will be more powerful (Lazar et al., 2017).

3.2.5 Quantitative data. To perform automatic data collection, this study used a Web application software which was explicitly created for the sole purpose of running this experiment. This tool presented participants with a series of tasks to be completed and registered data such as task completion time in minutes, produced image description by participants as a text, difficulty levels as the Likert-type items having rating scale from 1 to 4, and participants basic information—age group, gender, education nationality, email address, profession etc. The further section, *see section 4.1*, will explain experimental set up, tool, and assignments in detail.

3.2.6 Data Management. In order to manage the data, this study used database approach. Log file was also there as an alternative method to store the data but as it might require additional tools for interpretation or to parse data in a certain format (Lazar et al., 2017), this study went with the relational database which does not require additional data parsing to give it the format helpful for the data analysis. In addition, carefully designed relational databases can be used to store each action of interest in one or more database tables, together with all other relevant information (Lazar et al., 2017).

3.2.7 Image type Selection. Three different types of images—graph, map, and general image— were selected in order to identify the differences in the effectiveness of

cues based on the image types while writing descriptions. Here, the general image represents photos having humans or other nonliving objects. The graph and map images were taken from the NCAM guidelines Webpage and the general image was taken from the online image dataset called the pascal-sentences provided in (<http://vision.cs.uiuc.edu/pascal-sentences/>) which is particularly used for the image description evaluation. The selected images in this study and the selection criteria are given below.

3.2.7.1 Main Image Selection. Here, the main image indicates the image for the participants to write text description. Since, in this study, the evaluation of the description was supposed to be done based on the NCAM guidelines, it was carefully considered if the following properties explained in the guidelines were available in the selected image.

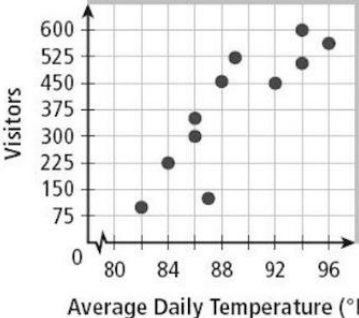

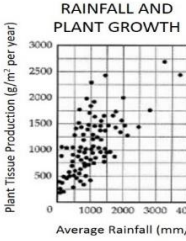



- In case of graph image, this study selected the scatter plot graph because it has been mentioned as one of the popular and difficult graphs to explain (NCAM, 2009). For the description of the scatter plot, the NCAM guidelines explain particularly about mentioning the title and axis labels; identifying image as a scatter plot and focusing the change of the concentration. Therefore, during the selection of the scatter plot image, the study considered if the selected image shows variation in the concentration; and has clear title and axis label. Furthermore, the study also considered if the axis has a well specified unit, for example, if the temperature is given in the axis then the unit should be specified in degree centigrade or degree Fahrenheit or any other units for the temperature but not only the numeric values as temperature.
- In case of a map, the NCAM guidelines explain specifically about central teaching point to determine if borders, region shapes, and bodies of water are important. Therefore, during the selection of the map, this study considered if the selected image provides some central teaching point which means if the map has some special region or border that it is intended to inform about. Furthermore, the NCAM guidelines also mention about the way how someone should explain about surrounding text and detailed caption available on the

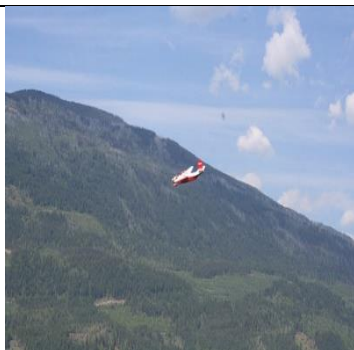
map. That is why this study also considered if the selected map has surrounding text.

- In case of a general image, the NCAM guidelines suggest that the description should be concise, objective, and grammatically correct. This study considered the image having many objects and a group of people showing different emotions. This is because this study wanted to see if the participants explain these contents according to the guidelines after having the hints.

3.2.7.2 Similar and Random cues Selection. While selecting the similar image as a similar cue, see *Table 3.1*, for the general image category, this study considered the object shown in the image as well as the context or surroundings of the object in the image. If the object and the context of that object is similar to the main image, then it was considered as a similar cue to the main image otherwise considered as the random cues. Likewise, for the graph and map, the images from the graph and map category was considered as a similar cue and out of this category was considered as a random cue.

SIMPLIFYING AND IMPROVING IMAGE DESCRIPTION USING SAMPLE CUE

Selected main Image	Random cue	Similar cue
<p>Beach Visitors</p>  <p>Visitors</p> <p>Average Daily Temperature (°F)</p>	 <p>"A cyclist relaxes on a bench and gazes toward the ocean"</p> <p>" Man sits on bench and looks toward ocean with a bike at his side"</p> <p>" Person sitting on a bench taking a break from bike riding"</p> <p>" Someone sitting on a bench next to a bike looking out over the sea"</p> <p>" Woman sitting next to bicycle on a snowy bench facing the ocean."</p>	<p>The graph below shows the relationship between annual rainfall and plant tissue growth rates in an ecosystem.</p>  <p>RAINFALL AND PLANT GROWTH</p> <p>Plant Tissue Production (g/m² per year)</p> <p>Average Rainfall (mm/year)</p> <p>" The graph is a scatter plot, entitled 'Rainfall and Plant Growth'. The horizontal X axis shows Average Rainfall ranging from zero to four thousand, in units of millimeters per year, in increments of one thousand. The vertical Y axis shows Plant Tissue Production in units of grams per meter squared per year, ranging from zero to three thousand, in increments of five hundred. The graph has approximately 85 points scattered in a pattern beginning in the lower-left corner where Plant Tissue Production and Average Rainfall are the lowest. The pattern extends toward the upper-right corner where Plant Tissue Production and Average Rainfall are the highest. Most of points are concentrated in the lower-left corner and diminish in concentration as the pattern extends toward the upper-right corner. "</p>
<p>Fjords of Norway Locator Map</p> <p>Map ©2010 HQP hillmanwonders.com</p>  <p>NORWAY</p> <p>Oslo</p> <p>Geirangerfjord</p> <p>Sognefjord</p> <p>Bergen</p> <p>Naeroyfjord</p> <p>Svalbard Islands</p> <p>North Cape</p>	 <p>"Two gentlemen talking in front of propeller plane"</p> <p>" Two men are conversing next to a small airplane"</p> <p>" Two men talking in front of a plane"</p> <p>" Two men talking in front of a small plane"</p> <p>" Two men talk while standing next to a small passenger plane at an airport."</p>	 <p>" A map of North America shows the regions claimed by the English, the French, and the Spanish during the early days of colonization. The area that would become the state of Kentucky was claimed by the French. Arrows also indicate La Salle's 1679 and 1682 routes of exploration. "</p>



"A red and white plane flying on a sunny day"

" A small red and white plane is flying over a grassy hill"

" A white and red plane flying past a mountain"

" Red and white plane flying through the air "

" The red and white airplane is flying in front of the mountain."

"A group of elderly people poses around a dining table"

" A group of elderly people sitting around a dining table"

" A picture of elderly people waiting in front of a dinner table"

" Friends and family gather for an evening meal"

" Group of elderly people sitting around a table."

Table 3.1. Selected images with random and similar images.

3.2.8 Selection of the Guidelines. This study considered 14 different guideline statements taken from the NCAM guidelines (NCAM, 2009). The contextual image description—description based on the context of the use of the image—is out of the scope of this study, therefore, it considered only the guidelines that cover the accessible descriptive image description. The list of selected guidelines is given below:

3.2.8.1 Common Evaluation Guidelines (applicable for all image category).

1. The description should be succinct.
2. Color should not be specified unless it is significant.
3. The new concept or terms should not be introduced.
4. The description should be started with high level context and drilled down to details to enhance understanding.
5. The active verbs in the present tense should be used.

6. Spelling, grammar, and punctuation should be correct.
7. Symbols should be written out properly.
8. The description vocabulary should be added which adds meaning, for example, "map" instead of the image.

3.2.8.2 Specific Evaluation Guidelines for the Graph.

9. The title and axis labels should be provided.
10. The image should be identified as a scatter plot and be focused on the change of concentration.

3.2.8.3 Specific Evaluation Guidelines for the map.

11. The central teaching point should be focused to determine if borders, region shapes, and bodies of water are important.
12. The description should be organized using number lists and pull the most important information in the beginning.

3.2.8.4 Specific Evaluation Guidelines for the General Image.

13. Physical appearance and actions should be explained rather than emotions and possible intentions.
14. The material should not be interpreted or analyzed, instead, the readers should be allowed to form their own opinions.

3.2.9 Image Description Evaluation Criteria. This study performed a human judgement study for image description evaluation like (Elliott & Keller, 2013; Kuznetsova, Ordonez, Berg, Berg, & Choi, 2012) used in their study to complement automatic image description evaluation. The volunteer evaluators were selected from the academic field who had experience regarding image description and aware of NCAM guidelines (NCAM, 2009). The evaluators were asked to perform the judgements according to the criteria based on NCAM guidelines. In the judgement process, the evaluators were supposed to use the scale from 1 to 4. The low number was for if the criteria were not fulfilled, and the high number if the criteria were fulfilled. The set of criteria used in this study during image description evaluation is given below:

3.2.9.1 General Criteria Based on the General Guidelines.

Concise: give a high score if the description is succinct; color has not been specified unless it is significant, and the new concept or terms have not been introduced.

- Objective: give high score if physical appearance and actions have been explained rather than emotions and possible intentions; the material has not been interpreted or analyzed, instead the readers have been allowed to form their own opinions; and the uncomfortable and controversial content such as images related with politics, sex, religion etc. has not been omitted.
- General to Specific: give a high score if the description has been started with high-level context and been drilled down to details to enhance understanding; content has been segmented into the logical and digestible chunks.
- Grammaticality: give high score if the active verbs in the present tense has been used; spelling, grammar, and punctuation are correct; the abbreviations and symbols have been written out properly to ensure proper pronunciation by screen readers; the descriptive vocabulary has been added which adds meaning, for example, 'map' instead of 'image'. sometimes it is acceptable to break traditional grammatical rules for brevity and clarity but should be consistent in this practice.

3.2.9.2 Specific Criteria Based on the Specific Guidelines for Graph Image (Scatter plot). Give high score if the title and axis labels have been provided; the image has been identified as a scatter plot and been focused on the change of concentration.

Provide Ratings According to the Guidelines

Beach Visitors

Average Daily Temperature (°F)	Visitors
81	100
83	250
84	300
85	350
86	400
87	450
88	500
89	550
90	450
91	500
92	550
93	600
94	550

Description written by participant:
The graph image which shows the number of visitors according to the average daily temperature. According to this figure more the temperature, more will be the visitors in the beach.

The description is succinct
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

Color has not been specified unless it is significant
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The new concept or terms have not been introduced
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The description has been started with high level context and been drilled down to details to enhance understanding
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The active verbs in the present tense has been used
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

Spelling, grammar, and punctuation are correct
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

Symbols have been written out properly
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The description vocabulary has been added which adds meaning, for example, "map" instead of image
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The title and axis labels have been provided
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

The image has been identified as a scatter plot and been focused on the change of concentration
Strongly Disagree ☐ 1 ☐ 2 ☐ 3 ☐ 4 Strongly Agree

SAVE

Figure 3.2. An example webpage showing human judgement framework for image description used in this study.

3.2.9.3 Specific Criteria Based on the Specific Guidelines for map. Give high score if the central teaching point has been focused to determine if borders, region shapes, and bodies of water are important; the description has been organized using number/bulleted lists and has been pulled the most important information on the beginning of the description; has been described the general trends and been referred to large areas at once with the help of surrounding text, or a detailed caption.

3.2.9.4 Specific Criteria Based on the Specific Guidelines for General Image.

Give high score if the physical appearance and actions have been explained rather than emotions and possible intentions, in addition, the material has not been interpreted or analyzed, instead the readers have allowed forming their own opinions.

3.2.10 Statistical Analysis. This section includes the statistical tests, Friedman and Wilcoxon Signed Rank test, used in this study for the data analysis. In addition, it reflects a brief insight of the Likert scale analysis that was used to evaluate the image description in compliance with the NCAM guidelines.

3.2.10.1 Friedman test. It is the non-parametric alternative to the one-way ANOVA with repeated measures. It is useful to test for differences among groups when the dependent variable being measured is ordinal (Lund-Research, 2013). Furthermore, it can be used for continuous data that do not have fulfilled the necessary assumptions to run the one-way ANOVA with repeated measures (Lund-Research, 2013). For example, data which are not normally distributed i.e. data having marked deviations from normality.

In order to do this test, data needs to pass the following four assumptions (Lund-Research, 2013):

- Assumption 1: Within group design with three or more different occasions.
This study fulfilled this assumption by adopting the within group design during the experiment where the same group of participants was exposed to three different occasions.
- Assumption 2: the group is a random sample of the population.
This study fulfilled this assumption by selecting the participants who fulfilled the requirements in order to be a participant in this study, see *Section 3.2.11*, from all over the world.
- Assumption 3: the dependent variables should be measured at the ordinal or continuous level. Example for the ordinal variables include Likert scales and continuous variables incorporate revision time (measured in minutes/hours), intelligence (measured using IQ score), and so forth.
This study fulfilled this assumption also because the dependent variables in this study, see *Section 3.2.3*, are ordinal and continuous in nature.

- Assumption 4: samples do not necessarily be normally distributed.

This assumption was also fulfilled in case of time data since the data was not found normally distributed while conducting the normality test.

Reporting Results in Friedman test. It compares the mean ranks between the related groups and informs how the groups differed. It uses the test statistics value (*chi-square*), degrees of freedom (*df*), and the significance level, also known as asymptotic significance (*Asymp. Sig.*) which are all we need to report the results from the Friedman test (Lund-Research, 2013). However, it is crucial to note that it tells whether there are overall differences but does not pinpoint which groups specifically differ from each other. To do this, we need to run separate Wilcoxon signed-rank test which is explained in the further section.

3.2.10.2 Wilcoxon Signed-rank test. It is also a nonparametric test equivalent to the dependent t-test. When the data is not normally distributed and the dependent t-test is inappropriate, this test can be the alternative (Lund-Research, 2013).

In order to do this test, data needs to pass the following four assumptions (Lund-Research, 2013):

Assumption 1: same as *Assumption 3* in the previous test.

Assumption 2: same as *Assumption 1* in the previous test.

Reporting Results in Wilcoxon Signed-rank test. It looks, specifically, for the “Asymp. Sig. (2-tailed)” value, which is the p-value for the test. If the p-value is less than the significance level considered during the test, then it is considered that there is a significant difference between these two groups. The important thing to be noticed in this test is that the value for the significance level needs to be adjusted using a Bonferroni correction in case of multiple comparisons in order to avoid the type I error which means to conclude as there is a significant result when we should not. In order to calculate the adjusted p-value, we need to divide the p-value by the number of groups. For example, if we have p-value equals to 0.05, and comparing three different groups to each other, then the adjusted p-value will be $0.05/3$ i.e. 0.017 (Lund-Research, 2013).

Effect size. The effect size of this test is calculated by dividing the Z value by the square root of N. However, in this case, N=the number of observations over the two time points, not the number of cases (Pallant, 2007). The effect size is considered as

small, moderate, and large using criteria of .1=small effect, .3=medium effect, and .5=large effect (Cohen, 1988).

3.2.10.3 Likert Scale Analysis. Jamieson (2004) mentioned several arguments regarding if the Likert scale should be considered as an ordinal or interval data and concluded that it is better to consider it as an ordinal dataset because the intervals between values cannot be presumed equal. To support his argument, Jamieson (2004) mentioned an interesting reason, that is the average of fair and good cannot be fair-and-a-half good. Furthermore, in order to calculate the central tendency of ordinal data set, mode or median should be used instead of mean calculation. Likewise, for the appropriate inferential statistical analysis for ordinal data, a non-parametric statistical calculation such as chi-squared, Spearman's Rho, or the Mann-Whitney U-test should be used because the parametric test requires data of interval or ratio level (Jamieson, 2004).

Allen and Seaman (2007) conclude that the Likert scale data should not use parametric statistics but should rely on the ordinal nature of the data. In case of having individual questions that have Likert response options for the participants to answer, we need to analyze them as Likert-type items and in this case modes, medians, and frequencies are the appropriate statistical tools to use (Boone & Boone, 2012). On the other hand, if we have a series of Likert-type questions that when combined describe a personality trait or attitude, is called a Likert scale and we should use means and standard deviations to describe the scale (Boone & Boone, 2012).

3.2.11 Participants. Total 65 participants took part in this experiment. Detail information about the participants, selection process, and selection criteria are explained in the further sections.

3.2.11.1 Information About Participants. Among the total participants, 49% of them were aware of image accessibility and 51% participants were not. Likewise, 92% participants said they had no any previous experience of describing images for the visually impaired people and only 8% participants said they had. Most of them i.e. 80% participants were non native English speakers but 20% participants said they are native English speakers. Among all of the participants, around 48% said they are competent, 17% said average, and 35% said they are fluent in English. Furthermore, the

participants had different nationalities, see *Figure 3.3*. Similarly, around 94% participants had College/University education and only 6% participants were from high school level. Regarding professional background, the participants had different backgrounds, see *Figure 3.4*. However, most of them i.e. 63% were students. Likewise, 80% were male and 20% were female. Different age group involved in this experiment. Most of the participants i.e. 49% were in the age group (32-36) and the least number was for the age group (45-50) which was 1%.

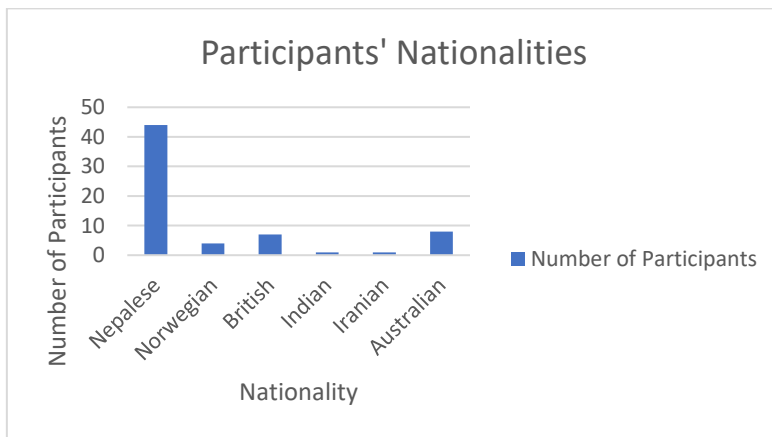


Figure 3.3. Several participants according to their nationalities.

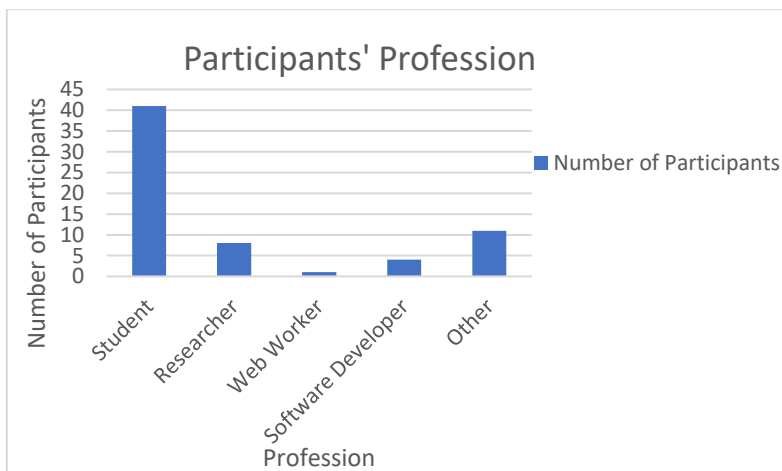


Figure 3.4. Several participants according to their professions.

3.2.11.2 Selection Process. First of all, this study put a post on the Facebook and other social networking sites about this research and requested for the participation. The post included the participant's criteria, see *Section 3.2.11.3*, and the motivational information, such as how an image description can help disabled people in the digital

society, to make someone participate in this work. Those who were interested and agreed the consent form were taken as the participants of this study.

3.2.11.3 Selection Criteria. Following are the criteria this study considered while selecting the participants:

- people having at least high school education who can read and write English and,
- who have experience of posting contents including non-textual contents such as image and video on the Web.

3.3 Pilot Study Before Actual Experiment

First, this study tried to figure out the accessibility issues on the Web experiment tool such as if the provided input fields were keyboard accessible, color contrast was sufficient, proper headings were used on the forms and other pages, enough instructions were provided to guide the participants, possible to go to the back page, proper form fields order were maintained, progress bar was available etc. These are basic guidelines defined in WCAG 2.0 (Ben Caldwell et al., 2008) in order to enhance the accessibility of a user interface.

To identify the probable bias on the actual data collection procedures, this study conducted the pilot studies with five participants. Each participant was asked to perform the task according to the instructions provided in a written and verbal form. In addition, 2 of the participants were observed by the researcher while performing the task to understand their meanings, viewpoints, values, and problems (Becker & Geer, 1957) towards the application user interface.

Through the pilot study, several potential biases were explored. Most of the participants complained about the image quality. According to them, some of the given images were not clearly visible and identifiable to the sufficient level. Likewise, the size of the example images was also not sufficient to understand the content. Furthermore, 2 of the participants were not satisfied with the position of the rating scale and 3 of them complained about the complexity of the graph and map in terms of writing the description about it. Likewise, during the observation session, the researcher noticed that they were confused about the position of the rating scale and pressed 'Ok' button

without assigning any ratings. In addition, they had to do zoom in to see the content of the image in case of the map.

This study updated the application in order to address the issues came out from the pilot study and made it ready for the real experiment. For this purpose, this study changed the size of the image to the bigger size. Likewise, this study replaced the previous map with the new, clear and simple, map. Furthermore, the graph was also replaced by a scatter plot graph with better visual quality.

4. Experiment and Results

In this chapter, the experimental setup and approach together with data findings and analysis reports are presented. *Section 4.1* includes several steps the participants were exposed during the experiment, in addition, describes an experimental tool which was developed explicitly to perform this experiment in detail. Results are incorporated in *Section 4.2* where line chart or bar graphs are presented to illustrate the percentage of the collected image descriptions based on different rating scales of the Likert type items evaluation framework. Graph descriptions are followed by the statistical analysis developed through the Friedman and Wilcoxon signed-rank tests, *see Section 3.2.10*, in order to find out if there exist any significant differences in the results that are shown on the bar or line graphs. Through these tests, it confirms whether the predefined null hypotheses of this study, *see Section 3.2.2*, are rejected.

4.1 Experimental Setup and Approach

This section describes how the online experiment was organized step by step. Furthermore, it gives information about the assignment participants had to perform. It also talks about several conditions/steps participants were exposed during this experiment.

A link was provided to the participants to the Web application created for this experiment. In this Web application, the first page contained the consent form, *see Figure 4.1*, to confirm that the participant was aware of what s(he) was going to do in this experiment and was ready to participate in this experiment as a voluntary task. The consent was written in a simple English language avoiding jargons so the participants with different background could read and understand it. The participants were supposed to check the checkbox to confirm that they were agreed with the information given in the consent form.

Oslo Metropolitan University
 Department of Computer Science
 Faculty of Technology, Art and Design

Consent Form

Title of Study: Evaluation of Supportive Hints for Writing Image Description

Investigator **Name:** Dhruha Dahal **Department:** Computer Science **Email:** s310279@stud.hioa.no

Introduction

This research is about writing the image/photo description. An image/photo description is a textual explanation about what the image/photo contains in it. You are being invited to participate in an online experiment where you will be writing several text descriptions for different images.

Description of the Study Procedures

If you agree to be a participant then you will be asked to do the following:

1. Pre-Questionnaires:

You need to complete a questionnaire based on the general information

2. Do Experiment:

In this experiment, there will be three rounds: first, second, and third round. You will be provided three different images in each round. The next round will start automatically when you finish writing for the current round. The overall experiment takes around 10 minutes.

Voluntary Participation:

Your participation in this study is completely voluntary.

Right to ask Questions and Report Concerns:

You have the right to ask questions about this research.

If you have any further questions about the study, at any time feel free to contact on the given information:

Name: Dhruha Dahal
Email: s310279@stud.hioa.no, dhruha.dahal03@gmail.com

Thesis Supervisor:

Name: Raju Shrestha, Associate Professor
Email: Raju.Shrestha@hioa.no

Confirmation and Consent

* ☐ I have read the above information and ready to participate in this experiment.

Figure 4.1. An online consent form.

After getting the agreement, the participants were directed to the next page, see Figure 4.2, where the participants were asked to enter the basic information such as age; gender; education; profession; nationality, if the participant is a native English speaker, level of English language (average, competent or fluent); previous experience of writing image description for someone who is visually impaired; and the email address. In addition, to understand if the participant was aware of image accessibility,

the options were provided to answer the question 'What is image accessibility?'. The number of options was increased from three to six to reduce the probability of false categorization. The participant who selected the right option was recognized as the participant who was aware of image accessibility.

Most of the field in the general information form were made compulsory and provided (*) sign in the red color. However, the field for the English language was optional because this study considered that some of the participants might hesitate to rate their English knowledge. Likewise, for the field of nationality, the additional validation was performed by checking if the participants entered a single letter or just some special characters. Furthermore, it was also checked if the email address entered was in a proper format.

Online Experiment
for
Image Description

Please Enter the Following Information

Your Age Group: ☒

Your Gender: ☒

Your Education: ☒

Your Profession: ☒

Are you a Native English Speaker?: ☒

Your English Language Level: ☒

Your Nationality: *

Have Previous Experience of Describing Images (for Visually Impaired) : ☒

What is an Image Accessibility? It is more About : ☒

Your Email: *

Figure 4.2. A Web form to enter participant's information.

To discourage the same participant perform multiple times, this study checked if the email address was already registered in the application. If so, the participant was

informed that this email has already been registered and not allowed to login again. In addition, by considering the possibility of sharing the login page provided in the registered email address, this study also checked user name and password whether the person having the specific username and password has performed the assignment already.

After completing the general information form, the participants had to check their email to get the link to the experiment login page and the necessary username and password.

4.1.2 Assignment. The participants were asked to write the text descriptions to each of the given images.

4.1.3 Conditions/Steps. There were three conditions/steps in this experiment. In the first step (named first round in the actual experiment page), the participants were provided three different images from three different categories—the graph, map, and general image—without any sample cue with proper text descriptions. Likewise, in the second step, the participants were provided the same three images as in the first step, but this time they had three different sample cues on the side of the each of the given images. In this second step, the sample cues were random i.e. not similar to the given images. However, they could get some idea of how a text description looks like for an image. Finally, in the third step, the participants got the similar sample cue to the given images they described in the first and second steps. The considered criteria in order to select these images have been explained in *Section 3.1.2*.

4.1.4 Tool for the Experiment. This study developed a Web application explicitly for this experiment using Java spring framework, jQuery, and MySQL relational database. It contains multiple pages which are discussed below.

4.1.4.1 First step. The first-step page, see in *Figure 4.3*, contains a header showing three different rounds. Below the header, there is a progress number bar. It helps the participants to understand how many images left in the current round. Below this progress bar, the image is provided in a sufficient size to look at it and understand the content before writing about it. Just below the given image, one text area is there, where participants are supposed to write some text description for the image. The text area field allows writing the short as well as a long description. Just beside the image box,

the difficulty rating scale is available. Using this rating, the participants are supposed to express their experience regarding how difficult it was to write the description for the given image. It is not possible to get the new image without writing the text description and without providing the difficulty rating. If somebody clicks the Ok button without completing the task, s(he) gets a pop-up message saying please complete the specific task first to go to the next step.

Figure 4.3. Main page after logging in.

4.1.4.2 Second step. The second-step page, see in Figure 4.4, contains an information statement below the progress bar, asking participants to look at the example before writing the description. In addition, the example image lies just beside the given image to ensure that the participants will at least look at it. The size of the example image is similar to the given image. Below the example image, a number of possible text descriptions are provided in a blue color making distinct among the other contents.

Furthermore, the rating scale is same as in the first step, but with the different placement, which is below the given image.

PREVIOUS ROUND

SECOND ROUND

NEXT ROUND

1

2

3

Please look at the example below before starting to write the description

Beach Visitors

Average Daily Temperature (°F)	Visitors
82	100
84	225
86	300
86	375
88	150
88	450
90	525
92	450
94	550
94	600
96	575
96	600

Example Image Description

"A cyclist relaxes on a bench and gazes toward the ocean"

"Man sits on bench and looks toward ocean with a bike at his side"

"Person sitting on a bench taking a break from bike riding"

"Someone sitting on a bench next to a bike looking out over the sea"

"Woman sitting next to bicycle on a snowy bench facing the ocean."

How difficult was it to write the description after having the hint with picture?

Not Difficult 1 2 3 4 Very Difficult

Ok

Figure 4.4. Second step page with a random example.

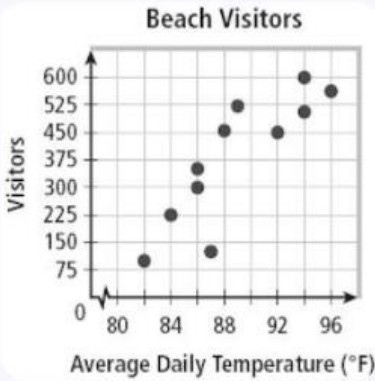
4.1.4.3 Third step. In the third or the final page, see in Figure 4.5, the information statement is presented just below the progress bar as in the second-step page, but this time it is in a different color than the previous round. The purpose of changing the color to orange from the black is to make the participants realize that the new step i.e. the third step has been started and, in this step, unlike the second step the example is

similar to the given image. The rest of the control elements are as it is like in the second step.

FIRST ROUND SECOND ROUND THIRD ROUND

1
2
3

Please look at the similar example below before starting to write the description



G

Example Image Description

The graph below shows the relationship between annual rainfall and plant tissue growth rates in an ecosystem.

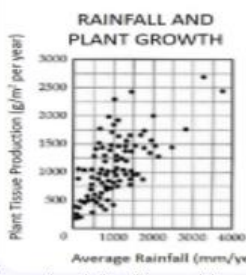


Image description: "The graph is a scatter plot, entitled ♦Rainfall and Plant Growth.♦ The horizontal X axis shows Average Rainfall ranging from zero to four thousand, in units of millimeters per year, in increments of one thousand. The vertical Y axis shows Plant Tissue Production in units of grams per meter squared per year, ranging from zero to three thousand, in increments of five hundred. The graph has approximately 85 points scattered in a pattern beginning in the lower-left corner where Plant Tissue Production and Average Rainfall are the lowest. The pattern extends toward the upper-right corner where Plant Tissue Production and Average Rainfall are the highest. The majority of points are concentrated in the lower-left corner and diminish in concentration as the pattern extends toward the upper-right corner."

How difficult was it to write the description after having the hint with similar picture?

Not Difficult 1 ☐ 2 ☐ 3 ☐ 4 ☐ Very Difficult

Ok

Figure 4.5. Third step page with a similar example.

4.2 Experimental Results

This section includes results from different combinations of the image types and sample cues. It also contains results based on the individual guidelines selected in this study.

4.2.1 Results with no cue, Random cue and Similar cue for the Common Guidelines.

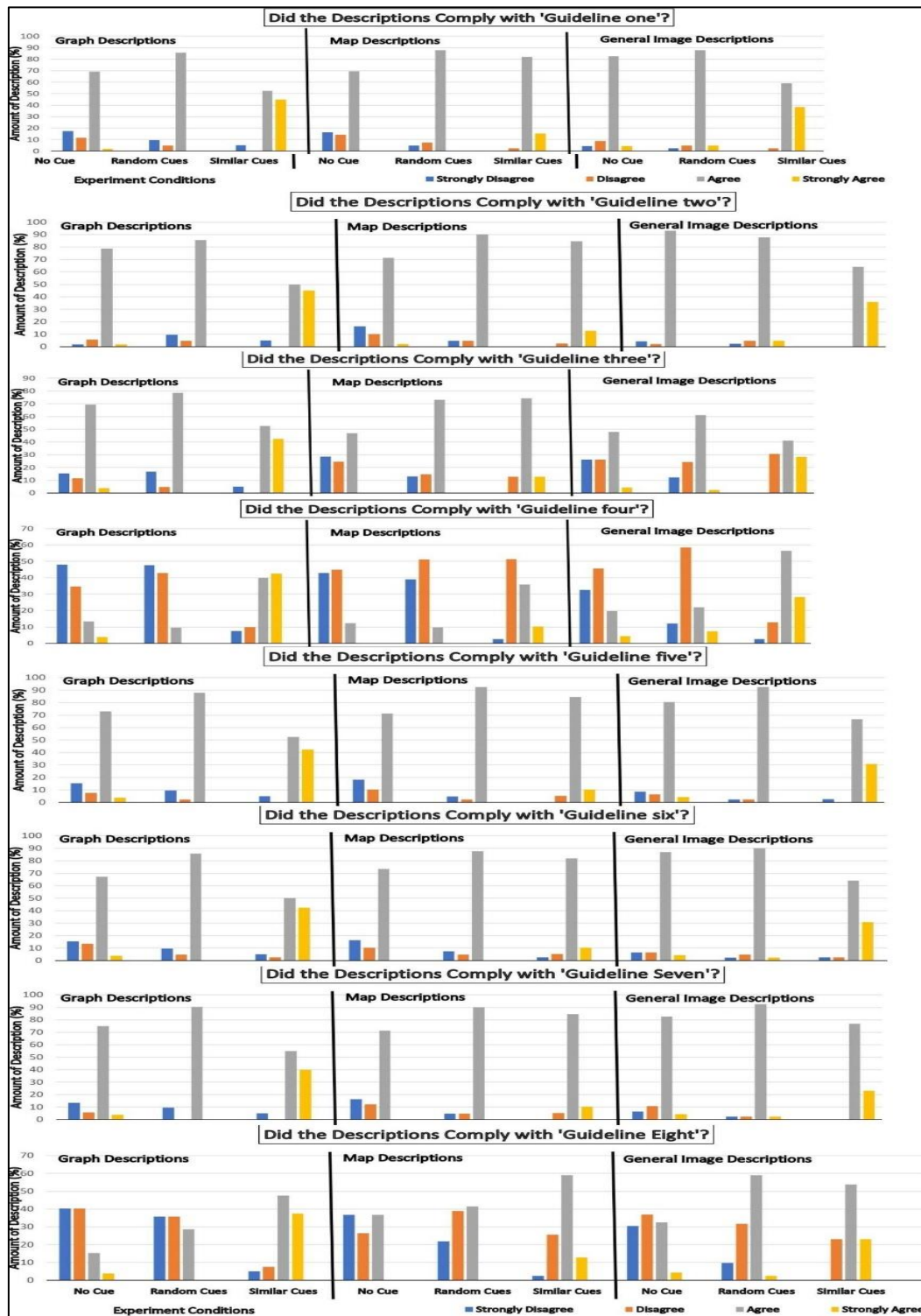


Figure 6. Graph showing the number of image descriptions in compliance with the common guidelines with no cue, random cue, and similar cue for all the three types of images.

Figure 4.6 shows the number of image descriptions in compliance with the common guidelines i.e. from guideline one to eight while having no, random, and similar cues for three different types of images. The further sub sections explain each of the guidelines.

- Guideline one. The above graph shows that most of the descriptions (around 70-80%) followed the guideline one—*description should be succinct*—without providing any cue for all kinds of images used in the study. However, with the similar cues, it was strongly agreed that the descriptions were written concisely compared to the random and no cue. Based on the image types, the number of strongly agreed brief descriptions is slightly higher in case of the graph and general image descriptions compared to the map descriptions in compliance with this guideline.
- Guideline two. The high percentage (70-90%) of the descriptions followed the guideline two—*color should not be mentioned unless it is significant*—without providing any cue for all kinds of images. However, like in guideline one, it was strongly agreed that the descriptions did not include any color descriptions with the similar cues. Based on the image types, the number of strongly agreed descriptions that did not explain the color is considerably higher for the graph and general image descriptions compared to the map descriptions in compliance with this guideline.
- Guideline three. A significant number of descriptions (40-65%) followed this guideline—*the new concept or terms should not be introduced*—without providing any cue for all kinds of images. However, the participants included new concepts or terms comparatively higher in the general image than in the map and graph descriptions. On the other hand, with the similar cues, participants did not include the new concepts or terms in all image categories except in the graph.
- Guideline four. Without any cues, almost 50% descriptions in all the image types did not follow the guideline four—*start with the high-level context and drill down to the details*. Likewise, the random cues also did not help them to get a clue in all the image categories since the strong disagreement is around equally higher. However, with the similar cues, the percentage of descriptions in compliance with this guideline increased by around 38% and 20% in the graph and general image

respectively. On the other hand, with the map, the percentage of descriptions that did not comply with this guideline is equally higher i.e. around 50% with no, random, and similar cues.

- Guideline five. The percentage of descriptions in compliance with this guideline— *should use active verbs in the present tense*— is also higher i.e. around 70-80% without providing any cue. However, after providing the random cues the percentage increased slightly i.e. by around 7% in the graph and general image descriptions and significantly i.e. by around 20% in the map descriptions. Likewise, with providing the similar cues, the strongly agreed percentage increased by around 10-40% in the three image types in compliance with this guideline.
- Guideline six. The above graph shows that there was almost no effect of cues to make the participants follow the guideline six— *correct grammar, punctuation, and spelling should be used*— in case of the general image descriptions. However, in the graph and map descriptions, the sample cues affected around 10% of the descriptions in compliance with this guideline.
- Guideline seven. Like in guideline six, there was almost no effect of cues in the general image descriptions to implement guideline seven— *symbol should be written out properly*. With the map, because of the random and similar cues, descriptions increased by 20% in compliance with this guideline. The effect of similar and random cues was almost similar in the graph and map descriptions.
- Guideline eight. Similarly, only 19% of descriptions followed the guideline eight— *better to add descriptive vocabulary*—without providing any cue for the graph. However, the percentage slightly increased by around 10% with the random cues and the percentage dramatically increased by 67% with the similar cues. For the map, the number of descriptions in compliance with this guideline was almost similar with no cue and random cues. However, the number increased by around 32% with the similar cues. With the general image, the number of descriptions in compliance with this guideline was again in the increasing fashion after providing the cues and the similar cues were comparatively more effective i.e. by around 14% compared to random cues.

4.2.2 Comparison of Effects of Similar and Random cues for Common Guidelines.

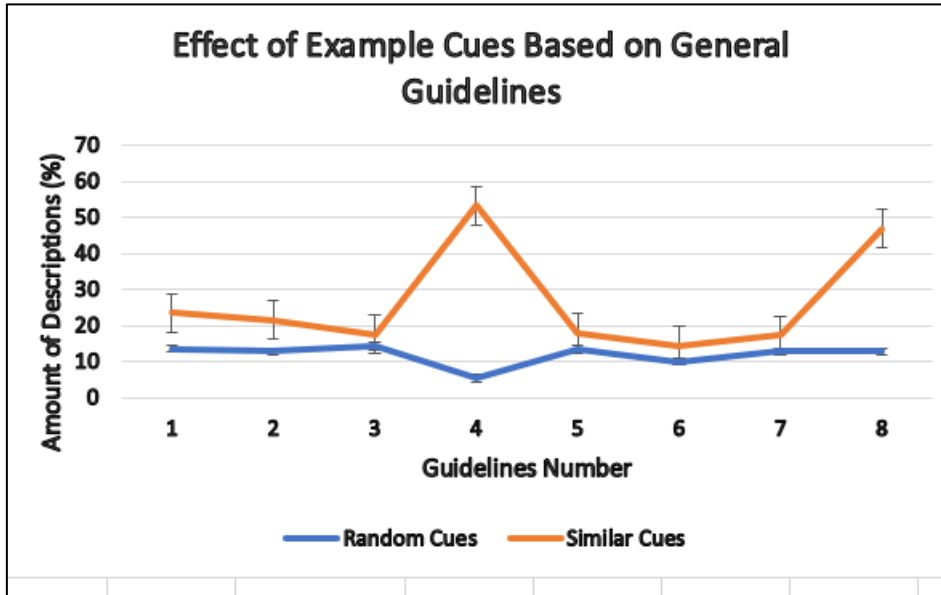


Figure 4.7. A line graph showing the average effect of sample cues based on the common guidelines for the graph, map, and general image.

The average difference in the number of descriptions between the random and similar cues in compliance with the common guidelines for three different types of images is lower i.e. around 2-10% except with the guideline four and eight which have significantly higher percentages i.e. around 47% and 36% respectively.

To make a conclusion, this study run the Friedman statistical test in order to reject or accept the third null hypothesis i.e. '*there is no significant difference in the effect of random and similar cues in compliance with the common guidelines*'.

Descriptive Statistics				
	N	Percentiles		
		25th	50th (Median)	75th
Random Cues	943	2.000	3.000	3.000
Similar Cues	943	3.000	3.000	4.000
No Cues	943	2.000	3.000	3.000

Friedman Test	
Ranks	
	Mean Rank
Random Cues	1.87
Similar Cues	2.41
No Cues	1.72

Test Statistics ^a	
N	943
Chi-Square	402.417
df	2
Asymp. Sig.	.000

a. Friedman Test

Figure 4.8. An analysis report based on the Friedman test for the effect of random and similar cues in compliance with the common guidelines.

Figure 4.8 shows the descriptive statistics, ranks and test statistics from the Friedman test performed in SPSS (IBM, 2018). The descriptive statistics table shows the median of ratings with no, random, and similar cues. Likewise, *Ranks* table shows the mean rank with three different cues. The final table, which is the most important, shows the number of observations as *N*, chi-square value, the degree of freedom *df* and a p value as Asymptotic Significance (Asymp. Sig.) which indicates evidence against the null hypothesis. Basically, we reject the null hypothesis with a small p value (typically less than or equal to 0.05) and accept the null hypothesis with a large p value (typically more than 0.05).

There was a statistically significant difference in compliance with the common guidelines depending on which type of cue (random or similar) was provided to the participants, $\chi^2(2) = 402.417$, $p = 0.000$. The third null hypothesis is rejected.

To examine how differently the types of cues affected the image descriptions, this study run separate *Wilcoxon signed-rank tests*, see Section 3.2.8.2, for the different combinations of related groups.

In *Table 4.1*, the Negative column implies that while calculating the ranks for the cue combination (a-b), the number of descriptions having higher ratings is greter with the cue b than with the cue a. The opposite is true for the Positive column. However, the Ties column represents the number of descriptions having equal ratings while having the cue a and b.

Ranks

	Negative	Positive	Ties
Random cue-No cue	178 ^a	288 ^b	477 ^c
Similar cue-No cue	92 ^d	506 ^e	345 ^f
Similar cue-Random cue	86 ^g	441 ^h	416 ⁱ

- a. No cue>Random cue
In 178 number of descriptions, the ratings with no cue were greater than with the random cue.
- b. Random cue > No cue
In 288 number of descriptions, the ratings with random cue were greater than with no cue.
- c. Random cue = No cue
In 477 number of descriptions, the ratings with random and no cue were equal.
- d. No cue>Similar cue
In 92 number of descriptions, the ratings with no cue were greater than the similar cue.
- e. Similar cue>No cue
In 506 number of descriptions, the ratings with similar cue were greater than no cue.
- f. Similar cue=No cue
In 345 number of descriptions, the ratings with similar and no cue were equal.
- g. Similar cue<Random cue
In 86 number of descriptions, the ratings with random cues were greater than the similar cue.
- h. Similar cue>Random cue
In 441 number of descriptions, the ratings with similar cue were greater than the random cue.
- i. Similar cue=Random cue
In 416 number of descriptions, the ratings with similar and random cue are equal.

Table 4.1. Rank table for similar-random cue combinations for the common guidelines.

Test Statistics^a

	Random cue-No cue	Similar cue-No cue	Similar cue-Random cue
Z	-5.282 ^b	-16.897 ^b	-14.538 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000
Effect size	-0.12 (low)	-0.39 (moderate)	-0.33 (moderate)

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

Table 4.2. Test statistics table for random-no, similar-no, and similar-random cue combinations for the common guidelines.

In no-random, no-similar, and random-similar cue combinations, the effectiveness of the cue in compliance with the common guidelines increased by low (-0.12), moderate (-0.383), and moderate (-0.42) size respectively, using Cohen (1988) criteria of .1=small effect, .3=medium effect, and .5= large effect.

4.2.3 Results with no cue, Random cue, and Similar cue for the Specific Guidelines.

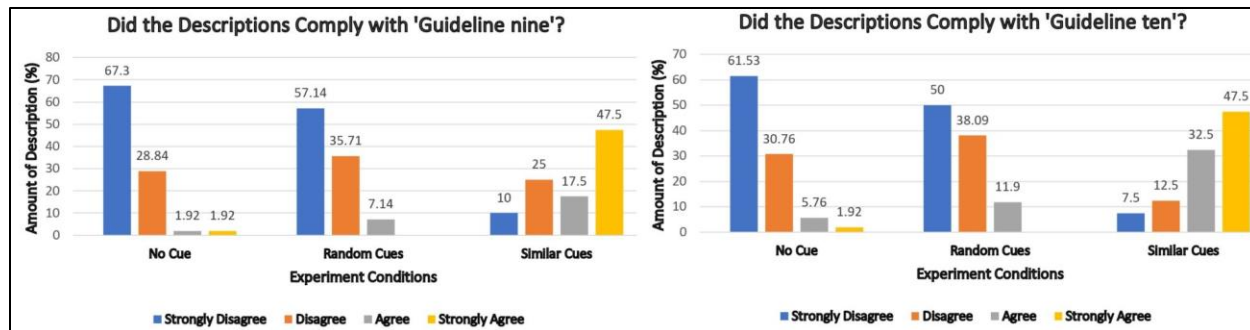


Figure 4.9. Bar graphs showing the descriptions in compliance with the specific guidelines for the graph while having no, random, and similar cues.

Figure 4.9 shows that most of the graph descriptions i.e. around 96% did not follow the guideline nine— *the title and axis labels should be provided*—without providing any cue. However, the percentage of descriptions in compliance with this guideline increased slightly by 4% with the random cue. Furthermore, the descriptions in compliance with this guideline increased significantly i.e. by around 55% after providing the similar cue.

Likewise, most of the graph descriptions, i.e. around 92%, did not follow the guideline ten— *the image should be identified as a scatter plot and be focused on the change of concentration*—without providing any cue. However, the percentage of descriptions in compliance with this guideline increased slightly by around 5% with the random cue. Furthermore, with the similar cues, the percentage increased dramatically i.e. by around 80%.

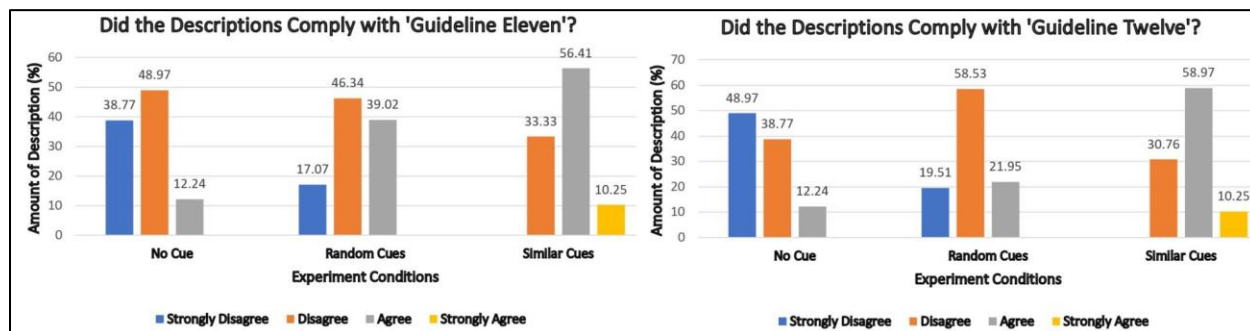


Figure 4.10. Bar graphs showing the number of descriptions in compliance with the specific guidelines for the map while having no, random, and similar cues.

Figure 4.10 shows that most of the map descriptions, i.e. around 88%, did not follow the guideline eleven— *the central teaching point should be focused to determine if borders, region shapes, and bodies of water are important*—without providing any cue. However, the percentage of descriptions in compliance with this guideline increased significantly by around 39 % with the random cue. Furthermore, the percentage increased sharply i.e. by 67% with the similar cue.

Similarly, most of the map descriptions i.e. around 88% did not follow the guideline twelve— *description should be organized using number lists and pull the most important information in the beginning*—without providing any cue. However, the percentage of the descriptions in compliance with this guideline increased significantly by around 22 % with the random cue. Furthermore, the percentage increased significantly i.e. by around 69% with the similar cue.

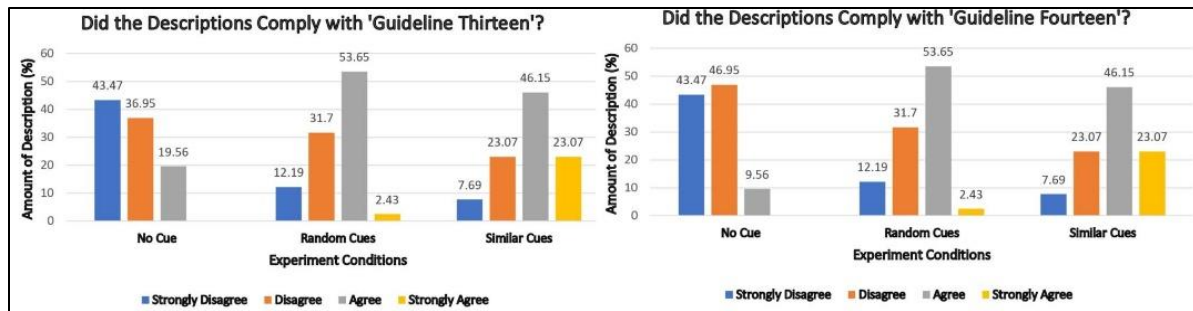


Figure 4.11. Bar graphs showing the number of descriptions in compliance with the specific guidelines for the general image while having no, random, and similar cues.

Figure 4.11 shows that most of the general image descriptions i.e. around 80% did not follow the guideline thirteen— *physical appearance and actions should be explained rather than emotions and possible intentions*—without providing any cue. However, the percentage of the descriptions in compliance with this guideline increased significantly by around 56 % with the random cue. Furthermore, the percentage increased considerably i.e. by around 69% with the similar cue.

Likewise, most of the general image descriptions i.e. around 89% did not follow the guideline fourteen— *the material should not be interpreted or analyzed, instead, the readers should be allowed to form their own opinions*—without providing any cue. However, the percentage of descriptions in compliance with this guideline increased significantly i.e. by around 46 % with the random cues. Furthermore, the percentage increased suddenly i.e. by 69% with the similar cue.

4.2.4 Comparison of Effects of Similar and Random cues for the Specific Guidelines.

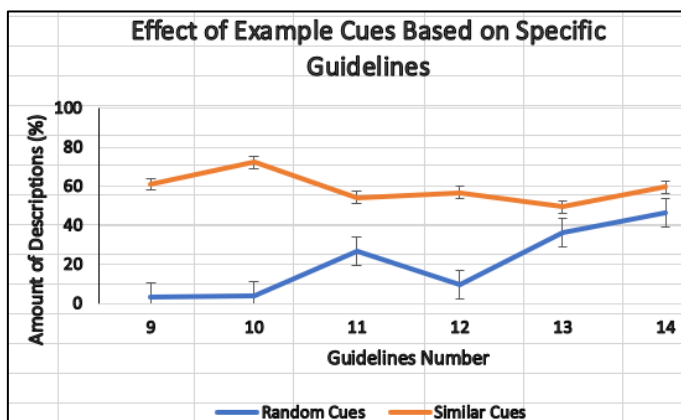


Figure 4.12. A line graph showing the average effect of example cues based on the specific guidelines for the graph, map, and general image.

The average number of descriptions in compliance with the specific guidelines—from the guideline nine to fourteen— is comparatively higher with the similar cue than the random cue. Specifically, for the guideline nine and ten, the difference in the number of descriptions with the random and similar cue in compliance with the guideline nine and ten is larger compared to other specific guidelines. The differences in the number of descriptions in compliance with the specific guidelines for the general image are lower while comparing with random and similar cues since the random cue also promoted the higher percentage around 35-50% of descriptions to comply with the guidelines thirteen and fourteen.

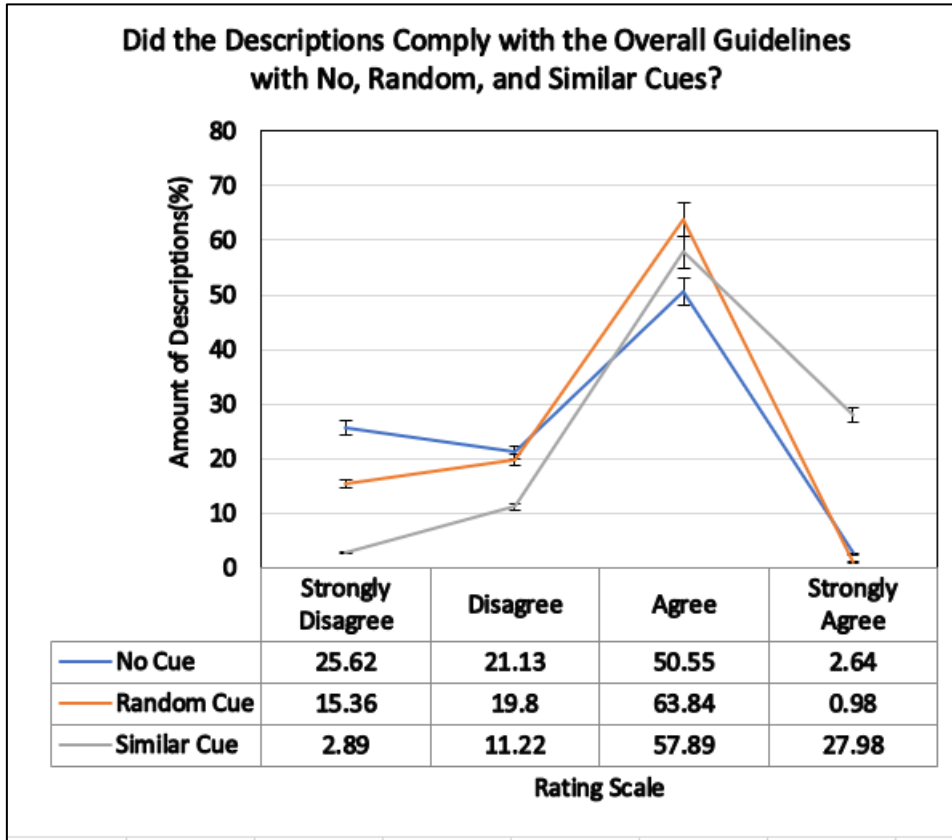
To make a conclusion, this study run a statistical test in order to reject or accept the second null hypothesis i.e. *'there is no significant difference in the effect of random and similar cues in compliance with the specific guidelines'*.

There was a statistically significant difference in compliance with the specific guidelines depending on which type of cue (random or similar) was provided to the participants, $\chi^2(2) = 177.506$, $p = 0.000$. The second null hypothesis is rejected.

To examine how differently the types of cues affected the image descriptions in compliance with the specific guidelines, this study run separate *Wilcoxon signed-rank tests*, see Section 3.2.10, on the different combinations of the related groups.

In no-random, no-similar, and random-similar cue combinations, the effectiveness of the cues in compliance with the specific guidelines increased by moderate (-0.40), high (-0.59), and moderate (-0.42) size respectively, using (Cohen, 1988) criteria of .1=small effect, .3=medium effect, and .5= large effect.

4.2.5 Results with no cue, Random cue, and Similar cue for Overall NCAM Guidelines.



Standard error 5%

Figure 4.13. A line graph showing the number of descriptions in compliance with overall guidelines while having no, random, and similar cues.

Figure 4.13 shows the number of descriptions in (%) in Y-axis in compliance with the rating scales from 'strongly disagree' to 'strongly agree' in X-axis. The three lines in different colors are for three different conditions—with no, random, and similar cues.

With no, random, and similar cues, the percentage of descriptions having a rating as 'agree' is comparatively higher than the other ratings. Almost 53 % (considering both 'agree' and 'strongly agree' percentage) of the descriptions complied with the overall guidelines without providing any cues. The percentage increased with the random cues by 12%. Likewise, the percentage increased significantly i.e. by 33% with the similar cues. In order to illustrate the statistical significance of the results with different cues, this study performed the Friedman statistical test, see Section 3.2.10.

There was a statistically significant difference in the compliance of guidelines depending on the types of cues (random or similar) provided to the participants, $\chi^2(2) = 544.655$, $p = 0.000$.

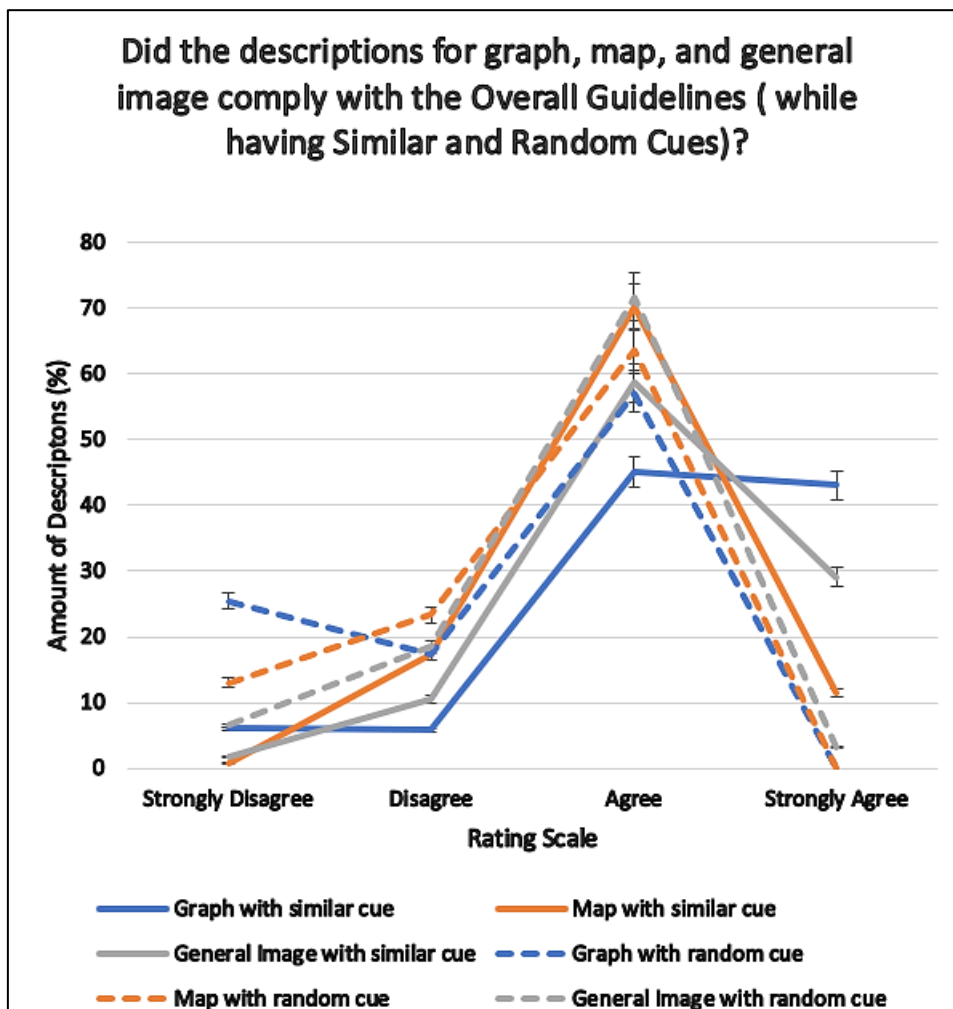
Since the test result is statistically significant with a p value less than 0.05, the first null hypothesis i.e. '*there is no significant effect of random and similar cues on image descriptions in compliance with the overall guidelines*' is rejected.

To examine how differently the types of cues affected the image descriptions in compliance with the overall guidelines, this study run the separate *Wilcoxon signed-rank tests*, see *Section 3.2.10*, on the related groups.

There was a statistically significant difference in the level of compliance with the guidelines depending on which type of cue was exposed to the participants, $\chi^2(2) = 544.655$, $p = 0.000$. Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, see *Section 3.2.10*, resulting in a significance level set at $p < 0.017$. There was a significant difference in the number of descriptions in compliance with the overall guidelines while having no cue and the random cue ($Z = -6.204$, $p = 0.000$), no cue and the similar cue ($Z = -19.769$, $p = 0.000$), and the random cues and the similar cue ($Z = -17.351$, $p = 0.000$).

In no-random, no-similar, and random-similar cue combinations, the effectiveness of the cues in compliance with the overall guidelines were increased by the small (-0.12), moderate (-0.414), and moderate (-0.363) size respectively, using (Cohen, 1988) criteria of .1=small effect, .3=medium effect, and .5= large effect.

4.2.6 Results with the Similar and Random cues Based on Image Types.



Standard error 5%

Figure 4.14. A line graph showing the number of descriptions in compliance with the overall guidelines based on the image types while having the random and similar cues.

Figure 4.14 shows the number of descriptions in (%) in Y-axis in compliance with the rating scales from 'strongly disagree' to 'strongly agree' in X-axis. The three solid lines in different colors are for three different image types with the similar cue. Likewise, the three dotted lines in different colors are with the random cue.

4.2.6.1 With Similar cue. With all the image types i.e. with the graph, map, and general image, the percentage of descriptions having a rating as 'agree' is comparatively higher than the other ratings. Almost 88 % (considering both 'agree' and 'strongly agree' percentage) of the graph descriptions complied with the overall guidelines while having the similar cues. The percentage decreased with the map by

6%. Likewise, the percentage increased slightly by 6% with the general image which is equal to the graph descriptions. In order to illustrate the statistical significance in the results with the different image types while having the similar cues, this study performed the Friedman statistical test, see *Section 3.2.10*.

There was a statistically significant difference in the effect of similar cues depending on the types of image (i.e. graph, map, and general), $\chi^2(2) = 73.383$, $p = 0.000$ with the level of significance 0.05.

To examine how differently the similar cues affected the image descriptions based on the image types, this study run separate *Wilcoxon signed-rank tests* on the different combinations of related groups.

Among the map and graph images, the effect of similar cues is significantly higher for the graph than the map since the negative value is higher than the positive value in the pair of Map-Graph images and $p=0.000$, see *Appendix A (3)*, but with low effect size i.e. -0.23, using (Cohen, 1988) criteria of .1=small effect, .3=medium effect, and .5= large effect.

Likewise, in case of the general image and graph, there is no any significant difference in the effect of similar cues since the p value (0.042) is greater than a Bonferroni correction applied significance level (0.017). Thus, the general image and graph were equally affected by the similar cues.

Among the general image and map, the effect of similar cues is significantly higher with the general image than the map since the positive value is higher than the negative value for the general image-map combination but with very low effect size i.e. -0.15, see *Appendix A (3)*.

4.2.6.2 With Random cue. With all the image types i.e. with the graph, map, and general image, the percentage of descriptions having a rating as ‘agree’ is comparatively higher than the other ratings. Almost 57 % (considering both ‘agree’ and ‘strongly agree’ percentage) of the graph descriptions complied with the overall guidelines while having the random cues. The percentage increased slightly with the map i.e. by 6%. Likewise, the percentage increased again slightly by 8% with the general image. In order to illustrate the statistical significance in the results with different

image types while having the random cues, this study performed the Friedman statistical test, see Section 3.2.10.

There was a statistically significant difference in the effect of random cues depending on the types of image (i.e. graph, map, and general), $\chi^2(2) = 39.837$, $p = 0.000$ with the level of significance 0.05. The fourth null hypothesis i.e. '*there is no significant difference in the effect of random and similar cues based on the image types in compliance with the overall guidelines*' is rejected.

To examine how differently was the image types affected by the random cues, this study run separate *Wilcoxon signed-rank tests* on the different combinations of related groups.

Among the map and graph images, unlikely with the similar cues, the effect of random cues is significantly higher for the map than the graph since the positive value is higher than the negative value and the p value (0.000) is less than a Bonferroni correction applied significance level (0.017) in the map-graph combination but with low effect size i.e. -0.13, see Appendix A (4).

Likewise, in case of the general image and graph, unlike with the similar cues, there is a significant difference in the effect of random cues since the positive value is higher than the negative value and the p value (0.000) is less than a Bonferroni correction applied significance level (0.017) in the general image-graph combination. Thus, the effect of random cues was slightly higher (-0.28) for the general image than the graph.

Among the general image and map, the effect of random cues is significantly higher for the general image than the map since the positive value is higher than the negative value for the general image-map combination and the p value (0.000) is less than a Bonferroni correction applied significance level (0.017). Thus, the effect of random cues was slightly higher (-0.15) for the general image than the map, see Appendix A (4).

4.2.7 Results for the Level of Difficulties.

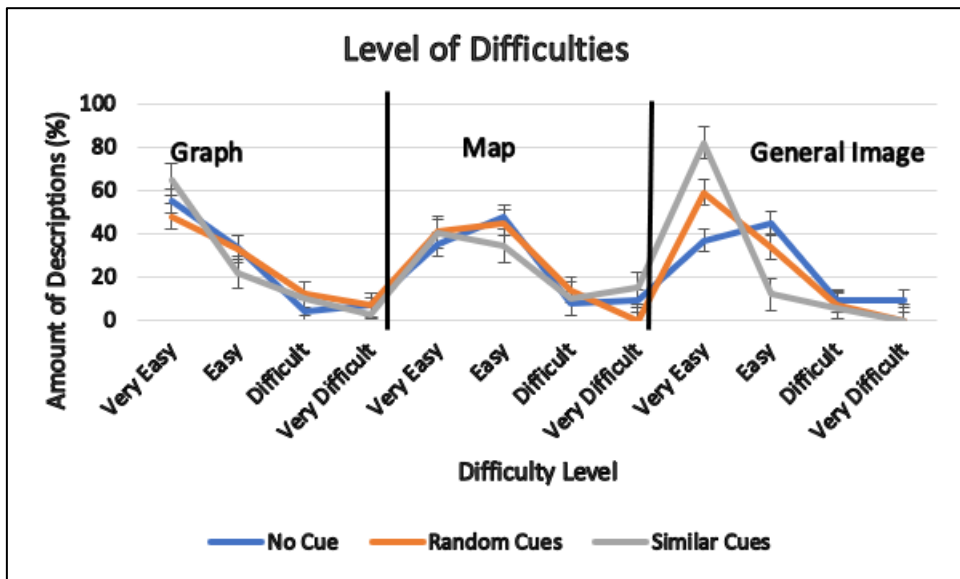


Figure 4.15. A line graph showing the level of difficulties experienced by the participants while writing the image descriptions in three different conditions.

Figure 4.15 shows that around 65% of the participants reported as easy to write the graph descriptions with the similar cues and the remaining mentioned some level of difficulties and very less i.e. around 3-4% mentioned as very difficult with the similar cues. On the other hand, around 48% of the participants reported as easy to write the graph descriptions with the random cues which is less than, by 2%, the participants who reported as easy without providing any cues.

Likewise, around 40% of the participants reported as easy to write the map descriptions with the similar and random cues. However, without any cues, around 25% participants mentioned difficulties while writing the map descriptions.

Furthermore, with the general image descriptions, around 82% of participants reported as easy to write the general image descriptions with the similar cues which is around 22% greater than the percentage with the random cues. In addition, no body mentioned as very difficult to write the general image descriptions with the random and similar cues. In the meantime, around 10% mentioned as very difficult writing the general image descriptions without any cues.

To make more strong conclusions and find out if the fifth null hypothesis i.e. *'there is no significant difference in the level of difficulties while writing image description with no*

cue, random cues, and similar cues' is rejected, this study run the Friedman statistical significance test, see *Section 3.2.10*.

There was a statistically significant difference in the level of difficulties to write image descriptions with no, random, and similar cues, $\chi^2(2) = 10.145$, $p = 0.006$, with the level of significance 0.05. The fifth null hypothesis is rejected.

To examine how differently was the level of difficulties affected by the types of sample cues, this study run separate *Wilcoxon signed-rank tests* on the different combinations of related groups.

In the random cue and no cue condition, the participants did not feel any significant difference in the difficulty level to write an image description with the p value 0.042, which is greater than a Bonferroni correction applied significance level (0.017), see *Appendix A (5)*.

Likewise, it was slightly difficult with the similar cue than providing no cue (-0.185) since the positive value is less than the negative value in the similar-no cue combinations, see *Appendix A (5)*.

Similarly, there is also no significant difference in the level of difficulties in the similar and random cue combination with the p value 0.235, which is greater than a Bonferroni correction applied significance level (0.017), see *Appendix A (5)*.

4.2.8 Results on Time Taken to Write Image Descriptions.

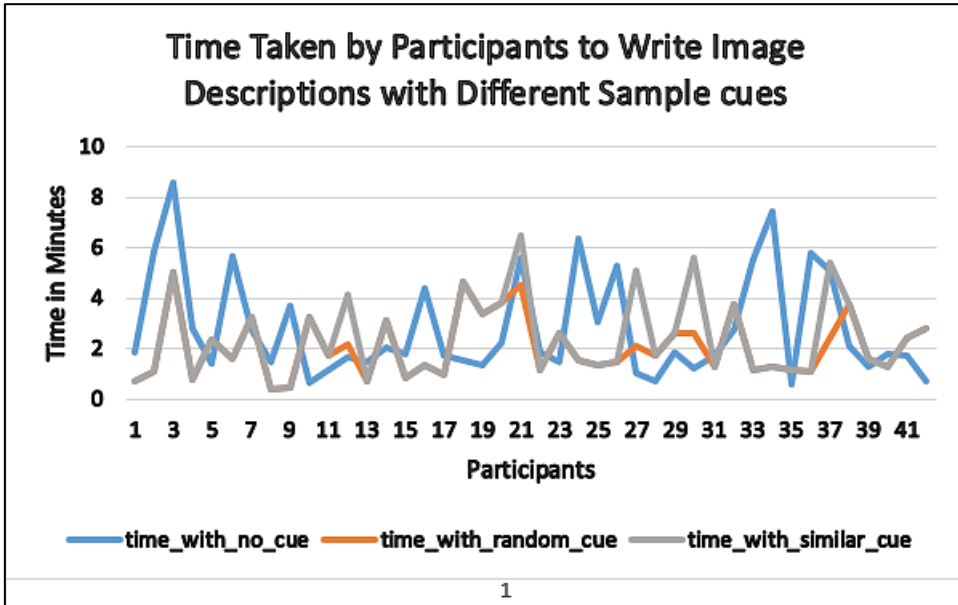


Figure 4.16. Line graph showing time taken by the participants to write the image descriptions with different sample cues.

The collected data for the time taken by the participants during this study did not fulfill the assumptions for one-way ANOVA with repeated measures. Therefore, this study conducted the Friedman test as an alternative to figure out if the sixth null hypothesis i.e. *'there is no significant difference in the time taken while writing image description with no cue, random cues, and similar cues'* is rejected.

The Friedman test showed that there are no any significant differences in time taken by the participants during the three experimental conditions i.e. no cue, random cue, and similar cue conditions, $\chi^2(2) = 4.014$, $p = 0.134$, with the level of significance 0.05. Hence, the sixth null hypothesis cannot be rejected.

5. Discussion

In this chapter, *Section 5.1* revisits the aim of the study and interprets the answers to the research questions. The findings from this study are discussed in relation to previous research in *Section 5.2*. In addition, *Section 5.3* reflects how useful the results of this study are. Not only that, *Section 5.4* includes the strong and weak aspects of this research. Likewise, several reasons why one can say the results of this study are significant are incorporated in *Section 5.5*. Possible generalization of the finds in *Section 5.6* and the potential criticisms of this study in *Section 5.7* are the other contents. Finally, *Section 5.8* wraps up this chapter with the ethical considerations.

5.1 Revisiting the aims of the Research

This research conducts an online experiment to explore the relationship between the sample cues and the quality of image descriptions in compliance with the NCAM guidelines. Based on the multiple guidelines taken from the NCAM guidelines as an evaluation framework, this research attempt to explain the effectiveness of the sample cues while writing the accessible image descriptions for three different image types—the graph, map, and general image.

The introduction to this thesis stated the first research question as this: *Can we simplify and improve quality (in compliance with the NCAM guidelines) of image descriptions for accessibility by providing sample cues?*

In order to answer the first research question, this study stated four hypotheses, see *Section 3.2.2*. The results from this study suggest that there is a significant effect of the similar and random cues on the image descriptions in compliance with the overall guidelines. In this way, the first null hypothesis is rejected. In addition, by comparing the effect of the similar and random sample cues, this study found that the similar sample cues are more effective than the random sample cues to make the descriptions comply with the overall guidelines. One of the reasons behind this result might be the provided similar cues in this experiment look very similar to the main image which might not always be possible to get in a real-life situation.

This study tried to explore if the different sample cues work differently for the different guidelines. The result from this study shows that the similar sample cues work significantly better than the random cues with a moderate effect size in compliance with the specific guidelines. This is how the second null hypothesis is rejected.

Furthermore, this study tried to figure out if the similar and random sample cues act differently with the common guidelines compare to the specific guidelines and the result shows that both cues are more effective for the specific guidelines than the common guidelines. However, the similar sample cues are more effective than the random cues with the common guidelines. More interestingly, the difference in the effectiveness of the random and similar sample cues are higher in case of the specific guidelines compare to the common guidelines. Thus, the third null hypothesis is also rejected.

Usually, it is more difficult to write image descriptions for the informative images such as the graph and map compare to other general images. Therefore, this study looked at how differently the sample cues work with the image types. The result of this study shows that there is a significant difference in the effect of the random and similar sample cues based on the image types in compliance with the overall guidelines. Among the map and graph, the similar sample cues worked almost equally with a low difference. However, the similar sample cues worked equally for the graph and general image with no difference in the effectiveness. Similarly, among the general image and map, the similar sample cues acted almost similar for both of them with a very low difference.

On the other hand, the random cues acted just opposite to the similar sample cues in case of the map and graph. That means the random sample cues were more helpful for the map than a graph. However, the difference is very low. Likewise, among the graph and general image, the random sample cues worked better for the general image with a low difference. Furthermore, like the similar sample cues, the random sample cues helped more for the general image than a map.

This study stated the second research question as this: *How are the difficulty level and time it takes for writing image description affected by a sample cue and different sample cues?*

In order to answer the second research question, this study stated two hypotheses, see *Section 3.2.2*. Based on the result of this study, the fifth null hypothesis is rejected but the sixth null hypothesis is accepted. The participants were asked to provide the Likert-type item rating based on the level of difficulty they felt while writing the descriptions with or without a sample cue for the graph, map, and general image. The result from this study shows that the participants found easier with the similar sample cues than without any cue. However, they did not find the random cue better than providing no cue in terms of writing difficulties.

Regarding the time taken by the participants, this study did not find any significant difference with and without the sample cues. It might be because of the online experiment where the participants are allowed to take a short or long break in between the experiment steps.

5.2 Major Differences and Similarities with Previous Research

The findings of this research support the previous work conducted by Morash et al. (2015) in which participants were provided templates based on the NCAM guidelines in order to write a graph description. The results from their research showed that the participants wrote the descriptions with more standardized in word use and content order while having the templates than having the set of guidelines. Similar to the results from Morash et al. (2015) study, this study illustrated that the sample cues were more helpful than not having any cue while writing the accessible image descriptions based on the NCAM guidelines. In addition, like the templates made less difficult to write the descriptions for the participants in Morash et al. (2015), the sample cues reduced the level of difficulties to describe the images in this study. Likewise, the time taken in the different experiment conditions/steps by the participants was not significantly different in both of the studies.

Morash et al. (2015) focused only on Science-Technology-Engineering-Mathematics (STEM) images. In contrast, this study incorporated not only the STEM images but also the general photos containing human and other non-living things such as a house, table etc. In this sense, the results from this study are more generalizable for the different image types compared to Morash et al. (2015). However, it seems necessary to conduct a comparative study between these two studies in order to conclude the one is better

than another in terms of producing more accessible and usable image descriptions, the writing complexities, and the required time duration.

The concept of providing sample cues while writing image descriptions does not interfere with the existing workflow. That is why this study is similar to the method suggested by Splendiani and Ribera (2014). Likewise, both methods are applicable in a Web context. Splendiani and Ribera (2014) argued that the cognitive load that is necessary to remember the information will be reduced due to the visual cue in a decision tree containing several captions. Similar to that, the method of providing the sample cues also reduce the cognitive load that is needed to memorize the way how an accessible image description should be written. However, the method suggested by Splendiani and Ribera (2014) is useful for the content authors to manage captions and make the most out of the captions. On the other hand, the method proposed in this study is useful not only for the users who are aware of image descriptions but also for those who do not have any idea about how to write an accessible image description.

5.3 Usefulness of the Study Results

The results from this study indicate that the sample images with standard text descriptions can be used in place of a set of guidelines in order to help Web users or workers writing accessible image descriptions while posting images on Web. Literature argued that the complexity of writing image description might be one reason to that many of images are without any proper text descriptions on Web. Therefore, with the real-time support, it reduces the complexity and can provide the idea of how one can write accessible image descriptions. It can be helpful not only for the novice Web workers who use different CMS such as WordPress in order for creating Web contents but also can motivate the normal users who use the internet to post their status and personal pictures in several social sites such as Facebook and Twitters. Several types of research are taking place in the field of image processing and the researchers seem to be interested in developing more reliable image matching algorithms that can identify semantically identical images (Ke, Sukthankar, Huston, Ke, & Sukthankar, 2004; Remondino et al., 2014). It shows a possibility to provide the sample images just before posting an image as a real-time support. Using this concept, we can facilitate the Web workers or the normal Web users to produce an accessible image description by

providing the similar sample images having accessible image descriptions on the Web content posting user interfaces.

Besides of making image description more accessible, there are other benefits of this sample cue method in a Web worker scenario. If the cue appeared just beside the image, it may work as a reminder for the person who is aware of the accessible images but does not usually remember to write text descriptions. Similarly, sometime the sample cue itself could be more suitable than the one which is going to be posted in the specific context. Hence, it might work as options provider, so the more suitable images can be posted with more accessible descriptions. Likewise, the Web workers might sometime get similar but have more details in the picture as a sample cue. In this case, the detailed information might be helpful to improve the content. Likewise, Russell et al., (2008) argue that the detailed information as an image label can be more helpful to retrieve the similar image from a database.

5.4 The Strong and weak Aspects of the Research

Following are the strength of this study:

- The significant number of participants having diverse background and nationalities participated in this study which makes the results more generalizable.
- In addition, because of the within group design, the tests were more powerful since the design provides an effective isolation of individual differences (Lazar et al., 2017).
- Furthermore, in order for the evaluation of the image descriptions, this study created a judgment group having six members who were aware of image accessibility, image description, and NCAM guidelines. Each member of the judgment group rated each of the descriptions based on the Likert-type items and the most common rating was selected through the mode calculation (most common item in a set of data). This is how the study reduced the biasness in the evaluation ratings.
- Accessible data collection online tool which was developed by following the WCAG guidelines ensured the possibility of the diverse user groups.

- Since the experiment was online and there was a high chance of multiple time inputs by the same participant, this study implemented two level authentication mechanism i.e. email based validation and username-password validation. The only way to get logged in to the tool for the second time was with a new email address. In this way, the participants were discouraged to write the descriptions multiple times.

The weaknesses of this study are listed below:

- Most of the experimental studies try hard for high reliability. One big challenge in HCI studies is that in contrast to the other sciences such as Physics, Chemistry, Biology etc., measurement of human behavior and social interaction are highly fluctuating, and therefore less replicable, in addition, it is hard to control confounding variable in experimental research (Lazar et al., 2017).
- As a weak part of this research, it took only three images in total to make the participants write descriptions. The first reason for taking such a small number was to avoid the longer experiment period that the participants do not entertain (Lazar et al., 2017). The next reason was the study did not have any computer-based algorithm to evaluate the descriptions in order to find out if they follow the NCAM guidelines, so the human judgement was necessary. Due to the time and resource limitation, it was not possible to take more images which cause a huge number of image descriptions and takes a lot of time and human resources. The results from this research would be more reliable if it could have managed to take more images. However, this study presented statistically significant results by conducting several statistical tests.
- As an evaluation framework, this study took the guidelines statement as it is from the NCAM guidelines. It might cause biasness because the guidelines are not very specific. However, this was mitigated by performing judgements by several people.
- This study considered only the quantitative data. It would be better if the study could have considered qualitative data also to support the results from the statistical calculation. Because of the time limitation, this study could not work

with the qualitative data. However, the study tried to mitigate this limitation by considering the level of difficulties the participants felt through the Likert-type item rating scale.

- This study did not perform the usability evaluation of the collected descriptions with the disabled people since they are the main user group of image description. However, this study considered the NCAM guidelines, developed based on multiple projects concerning the visual disabilities, see *Section 2.4.3*, as an evaluation framework.

5.5 Significance of Results

This study collected around five hundred image descriptions written for three different images in three different conditions with various sample cues. This is a statistically significant number to run a test to accept or reject the hypothesis. This study tried to explore the effectiveness of the sample cues in different levels based on the type of guidelines (general and specific) and image types (graph, map, and general image) with the different sample cues (random and similar).

To explore the level of difficulties experienced by the participants in three different conditions/steps, this study asked to fill the Likert-type item rating scale from one to four and performed the statistical test suitable for the ordinal data taken from 65 participants. Though this number gives the statistically significant results, the intense semi-structured interview with the participants about the usability experience could even be better for the significance of the results related to the user experience such as difficulty level.

Likewise, the time data could be even more significant if the study was conducted as a lab experiment instead of the Web experiment because in the lab experiment the environmental variables are more controllable than in the Web experiment (Lazar et. al., 2017).

5.6 Generalizing the Findings

This study focused on the descriptive image descriptions—which does not care about the image context but describes the image contents as neutrally as possible— rather than the contextual descriptions, which describe images based on the context they are used in. From that prospective, the results from this study cannot be generalized in the scenario where the context plays a significant role. However, for the descriptive image

descriptions, it covers several image types and diverse participants. That is why one can consider the results from this study generalizable to some degree. For example, the results from this study can be useful to support the normal Web users who are unaware of an accessible image description and do post pictures without proper text descriptions. On the other hand, it could also be supportive for the novice Web workers to gain an idea about an accessible image description while producing the Web content including images through several CMSs such as WordPress and Joomla.

5.7 The Potential Criticisms of the Research

In the current scenario in which the computer research field is moving towards the automatic generation of image descriptions, one might raise the question regarding how timely it is to talk about the human powered image descriptions. However, though the research is moving towards the automatic generation and has achieved some level of success to produce the shorter image captions, these captions are not enough to be an accessible image description. Today, people are still writing image descriptions by themselves while producing the professional Web contents such as news and other informative Websites. It reflects the usefulness of this research. In addition, a lot of people are posting pictures on Web without image descriptions. Computer algorithms are trying to understand those images and producing some text to explain. That means the current research trend is undermining the famous saying “prevention is better than cure” because in the current situation, people do not get any support describing images while posting the images on Web and are allowed to post images without any text descriptions. The aim of this research is to prevent the Web from the inaccessible images by encouraging people, at the right time, to write an accessible image description. This study believes that the person who is posting a picture is the most suitable one to describe that picture.

However, to have sample cues with the accessible image descriptions in a real life scenario might be an initial issue. To address this issue, the concept of a Web based tool for image annotation explained by Russell et al. (2008) might be useful.

5.8 Ethical Consideration

Walther (2002) explains human subjects research as a research in which there is any intervention or interaction with another person in order to gather information, or in which

information is recorded in such a way that a person can be identified directly or indirectly with it. Though the study asked the participants to write image descriptions, it did not ask the participants write identifiable data except the email address. However, the email addresses were also saved in a database in the encrypted form. Therefore, they were identifiable by no one including the researcher. In addition, they were used only for sending an experiment link to the participants and to avoid the multiple inputs for the same image by the same participant.

6. Conclusions & Future work

In this chapter, *Section 6.1* presents the conclusions that can be made from this study. Future works are included in *Section 6.2*.

6.1 Conclusions from the Research

This research presents how writing image description can be simplified and improve its quality in terms of accessibility by providing the sample cues.

The evaluation and analysis of the dataset illustrate different relationships among the type of guidelines, image types, and provided sample cues. In which, the similar sample cues were always better than the random cues and providing no cue. However, the difference in the effectiveness of the similar and random cue was less in case of the common guidelines compared to the specific guidelines. Similarly, the similar sample cues were comparatively more effective for the graph, general image, and the map respectively. But, the random cues were comparatively more effective for the general image, map and the graph respectively.

The dataset related to the level of difficulties collected from the participants and the automatically collected time data gave the ground to make some conclusions. With the similar sample cues, the participants felt less difficult, with a small difference, compared to having no cue. On the other hand, the random cues did not make any difference in the difficulty level compared to having no cue. Similarly, the participants took almost equal time with no, random, and similar cues.

6.2 Future work

This study suggests a comparative evaluation study of the template-based image description (Morash et al., 2015) and the cue-based image description discussed in this study as a future work. It would be possible to consider many observations resulting in more reliable results with the computer-based algorithm. Thus, this study reflects a need for a future study to develop an image description evaluation algorithm that can be used to evaluate the image descriptions in compliance with a guideline. Similarly, though this study performed an evaluation of the sample cues in compliance with the

NCAM guidelines, it still demands a concrete research to explore how should the sample cue be presented effectively on a user interface, so the diverse group of users can be benefitted while posting an image on Web.

References

- Aizpurua, A., Arrue, M., Harper, S., & Vigo, M. (2014). *Are users the gold standard for accessibility evaluation?* Paper presented at the Proceedings of the 11th Web for All Conference.
- Alampay, E. (2006). Beyond access to ICTs: Measuring capabilities in the information society. *International journal of education and development using ICT*, 2(3).
- Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, 40(7), 64.
- Awada, A., Issa, Y. B., Ghannam, C., Tekli, J., & Chbeir, R. (2012, 25-29 Nov. 2012). *Towards Digital Image Accessibility for Blind Users Via Vibrating Touch Screen: A Feasibility Test Protocol*. Paper presented at the 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems.
- Bavani, R., Azizah, J., & Yatim, N. F. M. (2010, 13-15 Dec. 2010). *A study on web experience among visually impaired users in Malaysia*. Paper presented at the 2010 International Conference on User Science and Engineering (i-USEr).
- Becker, H., & Geer, B. (1957). Participant observation and interviewing: A comparison. *Human organization*, 16(3), 28-32.
- BeMYEyes. (2017). Bringing Sight to the Blind and Visually Impaired. Retrieved from <http://bemeyes.com/>
- Ben Caldwell, T. R. D. C., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008). Web Content Accessibility Guidelines (WCAG) 2.0. Retrieved from <https://www.w3.org/TR/WCAG20/>
- BeneficentTechnology. (2017). Poet Image Description. Retrieved from <https://diagram.herokuapp.com/>
- Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., . . . White, S. (2010). *VizWiz: nearly real-time answers to visual questions*. Paper presented at the Proceedings of the 23rd annual ACM symposium on User interface software and technology.
- Bigham, J. P., Kaminsky, R. S., Ladner, R. E., Danielsson, O. M., & Hempton, G. L. (2006). *WebInSight:: making web images accessible*. Paper presented at the Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility.
- Bittar, T. J., do Amaral, L. A., Faria, F. B., & de Mattos Fortes, R. P. (2012). *Supporting the developer in an accessible edition of web communications: a study of five desktop tools*. Paper presented at the Proceedings of the Workshop on Information Systems and Design of Communication.
- Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of extension*, 50(2), 1-5.
- Brady, E. L., Zhong, Y., Morris, M. R., & Bigham, J. P. (2013). *Investigating the appropriateness of social network question asking as a resource for blind users*. Paper presented at the Proceedings of the 2013 conference on Computer supported cooperative work.

- Burgstahler, S. (2004). *Universal Design: Process, Principles, and Applications*. Retrieved from <https://www.washington.edu/doit/universal-design-process-principles-and-applications>
- Campbell, N. C., Elliott, A. M., Sharp, L., Ritchie, L. D., Cassidy, J., & Little, J. (2001). Rural and urban differences in stage at diagnosis of colorectal and lung cancers. *British Journal of Cancer*, 84(7), 910.
- Ching, C. (2004). Gender Governance: An Empowering Tool for E-Communities? *MEDIA ASIA-SINGAPORE*-, 31(2), 95.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* 2nd edn. In: Erlbaum Associates, Hillsdale.
- Connor, J. O. (2012). *Pro HTML5 accessibility*: Apress.
- Consortium, W. W. W. (2008). *Web Content Accessibility Guidelines (WCAG) 2.0*. Retrieved from <http://www.w3.org/TR/WCAG20/>
- Cooper, M., Sloan, D., Kelly, B., & Lewthwaite, S. (2012). *A challenge to web accessibility metrics and guidelines: putting people and processes first*. Paper presented at the Proceedings of the international cross-disciplinary conference on Web accessibility.
- Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*: Prentice Hall Upper Saddle River, NJ.
- Cundiff, C. (2015). Alt Text Bot. Retrieved from <https://connectability.devpost.com/submissions/37785-alt-text-bot>
- Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S., McCoy, K. F., & Chester, D. (2010). Interactive SIGHT: textual access to simple bar charts. *New Review of Hypermedia and Multimedia*, 16(3), 245-279. doi:10.1080/13614568.2010.534186
- Dikaiakos, M. D., Katsaros, D., Mehra, P., Pallis, G., & Vakali, A. (2009). Cloud computing: Distributed internet computing for IT and scientific research. *IEEE Internet computing*, 13(5).
- Dobransky, K., & Hargittai, E. (2006). The disability divide in Internet access and use. *Information, Communication & Society*, 9(3), 313-334.
- Doush, I. A., Pontelli, E., Son, T. C., Simon, D., & Ma, O. (2010). Multimodal Presentation of Two-Dimensional Charts: An Investigation Using Open Office XML and Microsoft Excel. *ACM Trans. Access. Comput.*, 3(2), 1-50. doi:10.1145/1857920.1857925
- Eggert, E., & Abou-Zahra, S. (2014, 23/04/2017). Web Accessibility Initiative. Retrieved from <https://www.w3.org/WAI/tutorials/images/>
- Elliott, D., & Keller, F. (2013). *Image description using visual dependency representations*. Paper presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.
- Facebook. (2017). Stats. Retrieved from <https://newsroom.fb.com/company-info/>
- Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., . . . Platt, J. C. (2015). *From captions to visual concepts and back*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). *Every picture tells a story: Generating sentences from images*. Paper presented at the European conference on computer vision.

- Feng, Y., & Lapata, M. (2013). Automatic Caption Generation for News Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4), 797-812. doi:10.1109/TPAMI.2012.118
- Ferres, L., Verkhogliad, P., Lindgaard, G., Boucher, L., Chretien, A., & Lachance, M. (2007). *Improving accessibility to statistical graphs: the iGraph-Lite system*. Paper presented at the Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility, Tempe, Arizona, USA.
- Francis, H., Al-Jumeily, D., & Lund, T. O. (2013, 16-18 Dec. 2013). *A Framework to Support E-Commerce Development for People with Visual Impairment*. Paper presented at the 2013 Sixth International Conference on Developments in eSystems Engineering.
- Freire, A. P., Russo, C. M., & Fortes, R. P. (2008). *A survey on the accessibility awareness of people involved in web development projects in Brazil*. Paper presented at the Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A).
- Fuglerud, K. S., & Sloan, D. (2013). *The link between inclusive design and innovation: some key elements*. Paper presented at the International Conference on Human-Computer Interaction.
- Goldstein, A., & O'connor, D. (2000). E-commerce for Development.
- Gomez, R., Hunt, P., & Lamoureux, E. (1999). Enchanted by telecentres: a critical look at universal access to information technologies for international development; paper presented at the conference New IT and Inequality, University of Maryland, Feb. 16-17, 1999.
- Gonçalves, R., Martins, J., & Branco, F. (2014). A Review on the Portuguese Enterprises Web Accessibility Levels – A Website Accessibility High Level Improvement Proposal. *Procedia Computer Science*, 27, 176-185. doi:<http://dx.doi.org/10.1016/j.procs.2014.02.021>
- Google. (2017). Search for pictures with Google Goggles. Retrieved from <https://support.google.com/websearch/answer/166331?>
- Harper, S., & Yesilada, Y. (2008). *Web accessibility: a foundation for research*: Springer Science & Business Media.
- Helbig, N., Gil-García, J. R., & Ferro, E. (2009). Understanding the complexity of electronic government: Implications from the digital divide literature. *Government Information Quarterly*, 26(1), 89-97.
- Helsper, E. (2008). *Digital inclusion: an analysis of social disadvantage and the information society*: Department for Communities and Local Government.
- Henry, S. L. (2005, 10/03/2017). Web Accessibility Initiative. Retrieved from <https://www.w3.org/WAI/intro/wcag.php>
- Henry, S. L., & McGee, L. (2016). W3C. Retrieved from <https://www.w3.org/standards/webdesign/accessibility>
- Hudson, S. E., & Mankoff, J. (2014). Concepts, Values, and Methods for Technical Human-Computer Interaction Research. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 69-93). New York, NY: Springer New York.
- IBM. (2018). IBM SPSS Software. Retrieved from <https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software>

- Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, 38(12), 1217-1218.
- Jayant, C., Ji, H., White, S., & Bigham, J. P. (2011). *Supporting blind photography*. Paper presented at the The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility.
- Karger, J., & Lazar, J. (2014). Ensuring that students with text-related disabilities have access to digital learning materials: a policy discussion. *Perspectives on Language and Literacy*, 40(1), 33.
- Karpathy, A., & Fei-Fei, L. (2015). *Deep visual-semantic alignments for generating image descriptions*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Ke, Y., Sukthankar, R., Huston, L., Ke, Y., & Sukthankar, R. (2004). *Efficient near-duplicate detection and sub-image retrieval*. Paper presented at the Acm Multimedia.
- Key, J. P. (1997). Research design in occupational education. *Oklahoma State University*.
- Kirkman, C. (2000). A model for the effective management of information and communications technology development in schools derived from six contrasting case studies. *Journal of Information Technology for Teacher Education*, 9(1), 37-52.
- Kohavi, R., & Longbotham, R. (2007). Online experiments: Lessons learned. *Computer*, 40(9).
- Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012). *Collective generation of natural image descriptions*. Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*: Morgan Kaufmann.
- Lazar, J., Goldstein, D. F., & Taylor, A. (2015). *Ensuring digital accessibility through process and policy*: Morgan Kaufmann.
- Leo, M., Medioni, G., Trivedi, M., Kanade, T., & Farinella, G. M. (2017). Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154, 1-15. doi:<https://doi.org/10.1016/j.cviu.2016.09.001>
- Lin, J., & Seepersad, C. C. (2007). *Empathic lead users: the effects of extraordinary user experiences on customer needs analysis and product redesign*. Paper presented at the ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.
- López, J. M., Pascual, A., Menduiña, C., & Granollers, T. (2012). *Methodology for identifying and solving accessibility related issues in web content management system environments*. Paper presented at the Proceedings of the International Cross-Disciplinary Conference on Web Accessibility.
- Lund-Research. (2013). Friedman Test in SPSS Statistics. Retrieved from <https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php>
- MacLeod-Clark, J. C., & Hockey, L. (1979). *Research for nursing: a guide for the enquiring nurse*: Wiley Heyden.

- Madhusudan, C. (2002). India's Hole in the Wall. *Key to Bridging the Digital Divide?* *TechKnowlogia*, 38-40.
- Maisel, J. L. (2010). *The State of the Science in Universal Design: Emerging Research and Developments*: Bentham Science Publishers.
- Matausch, K., Peböck, B., & Pühretmair, F. (2014). Accessible Web Content: A Noble Desire or a Need? *Procedia Computer Science*, 27, 312-317.
doi:<https://doi.org/10.1016/j.procs.2014.02.034>
- Moore, D., & McCabe, G. (1989). Introduction to the practice of statistics WH Freeman and Company. New York.
- Morash, V. S., Siu, Y. T., Miele, J. A., Hasty, L., & Landau, S. (2015). Guiding Novice Web Workers in Making Image Descriptions Using Templates. *Acm Transactions on Accessible Computing*, 7(4). doi:10.1145/2764916
- Morgan, G., Leech, N., & Barret, K. (2005). SPSS for intermediate statistics: Use and interpretation. In: New York: Lawrence Erlbaum Associates.
- Morris, M. R., Zolyomi, A., Yao, C., Bahram, S., Bigham, J. P., & Kane, S. K. (2016). *With most of it being pictures now, I rarely use it: Understanding Twitter's Evolving Accessibility to Blind Users*. Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.
- Mpofu, E., & Oakland, T. (2009). *Rehabilitation and health assessment: applying ICF guidelines*: Springer Publishing Company.
- NCAM. (2009). Guidelines for Describing STEM Images. Retrieved from http://ncam.wgbh.org/experience_learn/educational_media/stemdx/guidelines
- O'Farrell, C. (2001). *Information flows in rural and urban communities: Access, processes and people*. Paper presented at the UDRSA Conference.
- Organization, W. H. (2001). *International Classification of Functioning, Disability and Health: ICF*: World Health Organization.
- Pallant, J. (2007). SPSS survival manual, 3rd. *Edition*. McGrath Hill.
- Pascual, A., Ribera, M., & Granollers, T. (2012). *Perception of accessibility errors to raise awareness among web 2.0 users*. Paper presented at the Proceedings of the 13th International Conference on Interacción Persona-Ordenador.
- Power, C., Freire, A., Petrie, H., & Swallow, D. (2012). *Guidelines are only half of the story: accessibility problems encountered by blind users on the web*. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.
- Ramnath, K., Baker, S., Vanderwende, L., El-Saban, M., Sinha, S. N., Kannan, A., . . . Ramanan, D. (2014). *Autocaption: Automatic caption generation for personal photos*. Paper presented at the Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on.
- Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. *Psychological experiments on the Internet*, 89-117.
- Remondino, F., Spera, M. G., Nocerino, E., Menna, F., & Nex, F. (2014). State of the art in high density image matching. *The Photogrammetric Record*, 29(146), 144-166.
- Richardson, D., Ramirez, R., & Haq, M. (2000). Grameen Telecom's village phone programme in rural Bangladesh: a multi-media case study final report. *Consultado el*, 21.

- Rodr, S., #237, V, g., #225, zquez, & Lehmann, S. (2015). *Acrolinx: a controlled-language checker turned into an accessibility evaluation tool for image text alternatives*. Paper presented at the Proceedings of the 12th Web for All Conference, Florence, Italy.
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3), 157-173.
- Slatin, J. M. (2001). The art of ALT: toward a more accessible Web. *Computers and Composition*, 18(1), 73-81. doi:[https://doi.org/10.1016/S8755-4615\(00\)00049-9](https://doi.org/10.1016/S8755-4615(00)00049-9)
- Solomon, K. (2000). Disability divide. *The Industry Standard*. Retrieved June 2012. In.
- Splendiani, B., & Ribera, M. (2014). *How to textually describe images in medical academic publications*. Paper presented at the Proceedings of the XV International Conference on Human Computer Interaction.
- Stanford-University. (2015). Online Accessibility Program. Retrieved from <https://soap.stanford.edu/tips-and-tools/tips/image-descriptions>
- Statista. (2017a). Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017 (in millions). Retrieved from <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
- statista. (2017b). The Statistics Portal. Retrieved from <https://www.statista.com/statistics/657823/number-of-daily-active-instagram-users/>
- Story, M. F. (2001). Principles of universal design. *Universal design handbook*.
- Takagi, H., Harada, S., Sato, D., & Asakawa, C. (2013). *Lessons learned from crowd accessibility services*. Paper presented at the IFIP Conference on Human-Computer Interaction.
- TapTapSee. (2014). TapTapSee. Retrieved from <http://taptapseeapp.com/>
- Tariq, A., & Foroosh, H. (2017). A Context-Driven Extractive Framework for Generating Realistic Image Descriptions. *IEEE Transactions on Image Processing*, 26(2), 619-632. doi:10.1109/TIP.2016.2628585
- tilgjengelighetsloven, d.-o. (2008). diskriminerings- og tilgjengelighetsloven. Retrieved from <https://lovdata.no/dokument/LTI/lov/2008-06-20-42>
- Treviranus, J. (2008). Authoring tools. In *Web Accessibility* (pp. 127-138): Springer.
- Trewin, S., Cragun, B., Swart, C., Brezin, J., & Richards, J. (2010). *Accessibility challenges and tool features: an IBM Web developer perspective*. Paper presented at the Proceedings of the 2010 international cross disciplinary conference on web accessibility (W4A).
- Tumblr. (2017). Tumblr. Retrieved from <https://www.tumblr.com/>
- UN. (2006). Convention on the Rights of Persons with Disabilities (CRPD). Retrieved from <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html>
- Verhoest, P., & Cammaerts, B. (2002). Universal service: a tool for social and economic development?
- Von Ahn, L., & Dabbish, L. (2004). *Labeling images with a computer game*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

- W3C. (2005, 2015). Authoring Tool Accessibility Guidelines (ATAG) Overview. Retrieved from <http://www.w3.org/WAI/intro/atag.php#whatis>
- W3C. (2014a). Complex Images. Retrieved from <https://www.w3.org/WAI/tutorials/images/complex/>
- W3C. (2014b, 2017). Images Concepts. Retrieved from <https://www.w3.org/WAI/tutorials/images/>
- Walther, J. B. (2002). Research ethics in Internet-enabled research: Human subjects issues and methodological myopia. *Ethics and Information Technology*, 4(3), 205-216. doi:10.1023/a:1021368426115
- Wang, M., Sheng, Y., Liu, B., & Hua, X. S. (2010). In-Image Accessibility Indication. *IEEE Transactions on Multimedia*, 12(4), 330-336. doi:10.1109/TMM.2010.2046364
- Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide*: MIT press.
- Warschauer, M., & Newhart, V. A. (2016). Broadening our concepts of universal access. *Universal Access in the Information Society*, 15(2), 183-188.
- Wentz, B., Jaeger, P. T., & Lazar, J. (2011). Retrofitting accessibility: The legal inequality of after-the-fact online access for persons with disabilities in the United States. *First Monday*, 16(11).
- WordPress. (2017). WordPress Stats. Retrieved from <https://wordpress.com/activity/>
- Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017). *Automatic alt-text: Computer-generated image descriptions for blind users on a social network service*. Paper presented at the Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.
- Xiao, G., Xu, G., & Lu, J. (2010, 16-18 Oct. 2010). *iBrowse: Software for low vision to access Internet*. Paper presented at the 2010 3rd International Conference on Biomedical Engineering and Informatics.
- Yu, W., McAllister, G., Strain, P., Kuber, R., & Murphy, E. (2005). *Improving web accessibility using content-aware plug-ins*. Paper presented at the CHI '05 Extended Abstracts on Human Factors in Computing Systems, Portland, OR, USA.
- Zhong, Y., Lasecki, W. S., Brady, E., & Bigham, J. P. (2015). *Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users*. Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.

Appendixes

Appendix A: analysis reports

1. An analysis report for the effect of random and similar cues in the compliance of specific guidelines.

Descriptive Statistics				
	N	25th	Percentiles 50th (Median)	75th
Random Cues	157	1.000	2.000	3.000
Similar Cues	157	2.000	3.000	3.000
No Cues	157	1.000	1.000	2.000

Friedman Test	
Ranks	
	Mean Rank
Random Cues	1.98
Similar Cues	2.69
No Cues	1.33

Test Statistics ^a	
N	157
Chi-Square	177.506
df	2
Asymp. Sig.	.000

a. Friedman Test

Ranks

	Negative	Positive	Ties
Random-No	17 ^a	89 ^b	51 ^c
Similar-No	2 ^d	140 ^e	15 ^f
Similar-Random	18 ^g	97 ^h	42 ⁱ

a. No cue>Random Cues

b. Random Cues > No Cues

c. Random Cues = No cue

d. No cue>Similar Cues

e. Similar Cues>No cue

f. Similar Cues=No cue

g. Similar Cues<Random Cues

h. Similar Cues>Random Cues

i. Similar Cues=Random Cues

Test Statistics^a

	Random-No	Similar-No	Similar-Random
Z	-7.230 ^b	-10.399 ^b	-7.400 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

Effect size	-0.40 (moderate)	-0.59 (high)	-0.42 (moderate)
-------------	---------------------	--------------	------------------

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

2. An analysis report for the effect of no, random, and similar cues in the compliance of overall guidelines.

Descriptive Statistics				
	N	Percentiles		
		25th	50th (Median)	75th
No Cue	1140	1.000	3.000	3.000
Random Cues	1140	2.000	3.000	3.000
Similar Cues	1140	3.000	3.000	4.000

Friedman Test	
Ranks	
	Mean Rank
No Cue	1.69
Random Cues	1.86
Similar Cues	2.45

Test Statistics ^a	
N	1140
Chi-Square	544.655
df	2
Asymp. Sig.	.000

a. Friedman Test

Ranks

	Negative	Positive	Ties
Random-No	265 ^a	419 ^b	456 ^e
Similar-No	116 ^a	678 ^b	346 ^f
Similar-Random	97 ^c	567 ^d	476 ^g

a. Random Cues/Similar Cues < No Cue

b. Random Cues/Similar Cues > No Cue

c. Random Cues > Similar Cues

d. Random Cues < Similar Cues

e. Random Cues = No Cue

f. Similar Cues = No Cue

g. Similar Cues = Random Cues

Test Statistics^a

	Random-No	Similar-No	Similar-Random
Z	-6.204 ^b	-19.769 ^b	-17.351 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000
Effect size	-0.12 (low)	-0.414 (medium)	-0.363 (medium)

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

3. An analysis report for the effect of similar cues according to the image types

Descriptive Statistics				
	N	25th	Percentiles 50th (Median)	75th
Graph	351	3.000	3.000	4.000
Map	351	3.000	3.000	3.000
General	351	3.000	3.000	4.000

Friedman Test	
Ranks	
	Mean Rank
Graph	2.26
Map	1.73
General	2.02

Test Statistics ^a	
N	351
Chi-Square	73.328
df	2
Asymp. Sig.	.000

a. Friedman Test

Ranks

	Negative	Positive	Ties
Map-Graph	166 ^a	46 ^b	139 ^c
General Image-Graph	147 ^d	87 ^e	117 ^f
General Image-Map	66 ^g	137 ^h	148 ⁱ

- a. Graph>Map
b. Map> Graph
c. Map=Graph
d. Graph>General Image
e. General Image>Graph
f. Graph=General Image
g. Map>General Image
h. General Image>Map
i. General Image=Map

Test Statistics^a

	Map-Graph	General Image-Graph	General Image-Map
Z	-6.244 ^b	-2.036 ^b	-4.219 ^b
Asymp. Sig. (2-tailed)	.000	.042	.000
Effect size	-0.23 (low)	No effect	-0.15 (low)

- a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

4. An analysis report for the effect of random cues according to the image types.

Descriptive Statistics				
	N	25th	50th (Median)	75th
Graph	368	1.000	3.000	3.000
Map	368	2.000	3.000	3.000
General	368	2.250	3.000	3.000

Friedman Test	
Ranks	
	Mean Rank
Graph	1.83
Map	1.99
General	2.18

Test Statistics ^a	
N	368
Chi-Square	39.837
df	2
Asymp. Sig.	.000

a. Friedman Test

Ranks

	Negative	Positive	Ties
Map-Graph	82 ^a	127 ^b	159 ^c
General Image-Graph	53 ^d	133 ^e	182 ^f
General Image-Map	66 ^g	120 ^h	182 ⁱ

- a. Graph>Map
 b. Map> Graph
 c. Map=Graph
 d. Graph>General Image
 e. General Image>Graph
 f. Graph=General Image
 g. Map>General Image
 h. General Image>Map
 i. General Image=Map

Test Statistics^a

	Map-Graph	General Image-Graph	General Image-Map
Z	-3.467 ^b	-7.851 ^b	-4.203 ^b
Asymp. Sig. (2-tailed)	.001	.000	.000
Effect size	-0.13 (low)	-0.28 (low)	-0.15 (low)

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

5. An analysis report for the difficulty level the participants experienced to write image descriptions in three different experiment conditions.

Descriptive Statistics				
	N	25th	50th (Median)	75th
No Cue	118	1.000	2.000	2.000
Random Cue	118	1.000	2.000	2.000
Similar Cue	118	1.000	1.000	2.000

Friedman Test	
Ranks	
	Mean Rank
No Cue	2.17
Random Cue	2.01
Similar Cue	1.82

Test Statistics ^a	
N	118
Chi-Square	10.145
df	2
Asymp. Sig.	.006

a. Friedman Test

Ranks

	Negative	Positive	Ties
Random-No	44 ^a	32 ^b	42 ^c
Similar-No	55 ^d	27 ^e	36 ^f
Similar-Random	43 ^g	29 ^h	46 ⁱ

- a. No>Random
 b. Random> No
 c. Random=No
 d. No>Similar
 e. Similar>No
 f. Similar=No
 g. Random>Similar
 h. Similar>Random
 i. Similar=Random

Test Statistics^a

	Random-No	Similar-No	Similar-Random
Z	-2.034 ^b	-2.842 ^b	-1.189 ^b
Asymp. Sig. (2-tailed)	.042	.004	.235
Effect size	No Significant Difference	-0.185 (low difference)	No Significant Difference

- a. Wilcoxon Signed Ranks Test
 b. Based on negative ranks.

6. An analysis report for the time taken by the participants while writing image descriptions in three different experiment conditions.

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	25th	Percentiles 50th (Median)	75th
No Cue	120	1.794	1.6265	.2	8.6	.678	1.309	2.308
Random Cue	120	1.373	.9934	.2	5.0	.721	1.099	1.754
Similar Cue	120	1.807	1.4694	.2	6.5	.723	1.250	2.625

Friedman Test	
Ranks	
	Mean Rank
No Cue	1.97
Random Cue	1.90
Similar Cue	2.14

Test Statistics ^a	
N	120
Chi-Square	4.014
df	2
Asymp. Sig.	.134

a. Friedman Test

Appendix B: image description examples written by the participants

The following table shows a couple of representative image description examples written by the participants during this study.

	Example 1			Example 2		
Graph	No sample	Random sample	Similar sample	No sample	Random sample	Similar sample
	Number of beach visitors with respect to the average daily temperature.	"More the daily temperature more the beach visitors"	"This graph is a scatter plot of daily visitors in a beach, which shows the increase of visitors in the beach with respect to increase in daily temperatures".	The image shows the number of beach visitors in relation to daily temperature. The image shows that higher the temperature, higher is the number of people visiting beach. For example, when the temperature is 82 F, 75 people visited the beach. when the temperature is 94, the number of people is about 525 and 600	Number of beach visitors against average daily temperature. At 96 F, about 575 beach visitors.	The image is a scatter plot showing number of beach visitors at different temperatures. The number of visitors ranged from 75 to 600 in a vertical axis. The horizontal axis shows average daily temperature at increment of 4F. The number of people visiting beach is highest at 94F while lowest at 82F
Map	Locations of some of the Fjords in Norway	"More Fjords are located on west side of Norway"	" A map of Norway with the locations of major Fjords"	The image shows different fjords in Norway map. Up in the north there is Svalbard islands while down around Bergen, there are Geiragerfjord, Sognefjord, Naeroyfjord.	Map of Norway showing fjords in its different parts	The map shows locations of fjords in Norway. The image for example, shows that Sognefjord is nearby Bergen
General image	Picture of happy faces taken before the dinner:-)	"Finally Dinner Time - Looks delicious "	"It's a dinner time - Cheers !". Alternative caption: "A picture of family around the dinner table"	The image shows two ladies and a man having dinner perhaps. Though there are three people, there	two lady and a man having dinner in a place where a painting is hanging in the wall	A group of people posing for photograph around a dining table

SIMPLIFYING AND IMPROVING IMAGE DESCRIPTION USING SAMPLE CUE

				are four plates, perhaps the image is taken by their friend. The wall has a painting hanging on it. I can see a bottle of wine perhaps and food in the table		
--	--	--	--	--	--	--

Appendix C: relational database used in this study

