The Effect of Online Questionnaire Format on Website Assessment

Peter Moor

p.j.moor@student.utwente.nl

ABSTRACT

One way to evaluate websites is by using online questionnaires. Both websites and questionnaires are usually designed with respect to common guidelines.

The current study tried to find what effect the format of a questionnaire has on the assessment of a website, by comparing two questionnaires and two websites. One questionnaire and one website were better than the other ones, according to common guidelines.

Four conclusions have been drawn from the results. The most important one is that the effect of the questionnaire format is very small or does not exist at all.

Keywords

website evaluation, online questionnaire, website guidelines

1. INTRODUCTION

1.1 Using online questionnaires for website assessment

In the past decades, the use of internet has grown rapidly, making webpages a more common user interface each year. Just like other computer systems, websites must be designed to have good usability and user satisfaction. Guidelines for good website design have been put forward [Van98, Res].

Also, evaluation of websites is important [Alp99]. This can be done in several ways, for example with an online questionnaire. Research has been done on how to compose and evaluate questionnaires [Alp99, Tul04]. Since online questionnaires are realized as webpages, webpage guidelines apply to these questionnaires as well [Lum05]. Just like webpages in general, online questionnaires are supposed to be designed professionally and have good usability.

From now on, the words "professional" and "unprofessional" will be used to discriminate between websites or online questionnaires that respect the format guidelines (e.g. about used colors, fonts and navigation aids) more or less. In this text, these terms are only relative and are only meant to discriminate between websites or questionnaires, implying nothing about absolute professionality.

1.2 The effect of the questionnaire format

Professionality of a questionnaire seems important, for the word "professional" suggests neutrality. A questionnaire should be neutral because its design must not influence the results of the website assessment it is used for.

Can a questionnaire really be neutral? Perhaps participants in online website assessment are always influenced in one way or another. The professionality of the questionnaire might not simply guarantee professionality, but evoke a bias. For example, users might take the questionnaire format as the criterium for the website assessment, so they appreciate a website more when it looks equally professional.

Another possible effect of questionnaires is that they evoke positive or negative affect in participants, which they incorrectly attribute to the websites they are rating. According to the psychological attribution-of-arousal theory [Sch62, Gle04], people sometimes attribute their emotions to the wrong sources. This might cause people to rate websites higher with more professional questionnaires.

These and other effects are not implausible. The current study was an attempt to find out whether one of these effects could indeed be found. To do this, two questionnaires and two websites have been compared in a website assessment survey.

The remainder of this paper describes the method of the study that has been conducted, the acquired results, the conclusions that can be drawn from these results, and a discussion of the implications that this study has for both future research and online questionnaire design.

2. METHOD

To find the effect of online questionnaire format on website assessment, an online website assessment survey was conducted.

The first thing that was done for this study, was choosing which guidelines would be used to discriminate between more and less professional websites or questionnaires.

Two different questionnaires were used for the survey, having the same questions but different formats. One of them was much more professional than the other one, according to the chosen guidelines. Each participant had to finish one of these questionnaires, which was allocated randomly.

All participants were asked to assess the same two websites, although the order of these websites was randomized. These websites had the same subject, but differed significantly in professionality.

By comparing the ratings for the different websites and acquired with the different questionnaires, possible effects of the questionnaire format could be detected.

The remainder of this chapter describes how the guidelines were selected, which websites and questionnaires have been used, and how participants were found.

2.1 Guidelines

There is not a single set of guidelines that is used by everybody, although most research on guidelines seems to agree on the fact that accessibility is a very important concept [e.g. Aba04, Hac05, Zap05]. [Nie05] explains that usability is the really

important concept and accessibility is necessary for achieving this. But even this concept is not agreed on by many website designers, who regard accessibility as a constraint for their creativity [Pet04].

Attempts have been made to validate existing guidelines and put them together in a single collection [Van98, Res]. The guidelines used for this study are those stated by [Res]. These guidelines come from different sources and an indication of the strength of their evidence is given.

From the 55 guidelines, 32 have been selected because they could easily be checked, were relevant for the websites used and were clearly different from other guidelines. Small description of the guidelines can be found in Attachment A.

These guidelines have been used both to choose a professional and an unprofessional website and to create a professional and an unprofessional questionnaire.

The unprofessional website was selected for disrespecting many guidelines and thus having bad usability, but not being too inaccessible. Extreme inaccessibility would make it impossible for people to gather any information. For the same reason, the unprofessional questionnaire was created in a way disrespecting guidelines without being completely inaccessible.

2.2 Websites
Two websites about free software have been chosen, almost equal in number of software titles listed and number of software categories. The advantage of two websites equal in subject and size is that they are comparable. The specific subject of free software has been chosen for the practical reason that participants interested in this subject could easily be found.

Several websites have been rated on the number of guidelines supported or violated, and to what extent. (Some guidelines are "almost" supported, while others are severely violated.)

One website [Onl] violates about 8 of the guidelines, most of them only slightly. The website has a clear, consistent format with supportive navigation aids. The guidelines violated most seriously are probably the fact that links do not have the standard link format (underlined, blue, purple when visited) and the fact that some pages are a little large and should better have been split into several shorter pages. From now on, this website will be referred to as the professional website.

The other website [Abs] violates about 13 of the guidelines, some of them quite seriously. Fonts, titles, alignment and navigation links are not consistent, the home page being quite different from the other pages. Fonts are very small on most pages (although relative to the brower's settings), there is no clear distinction between internal and external links, supportive navigation aids are missing and images do not have alternative texts. Also, there are several less serious guideline violations. From now on, this website is called the unprofessional website.

Attachment A shows which guidelines have been respected or violated by these two websites.

2.3 Questionnaires
There are several questionnaires for assessing the usability of computer systems or websites, some of which have been compared in [Tul04].

In this study, the System Usability Scale (or SUS, [Bro96]) has been used, for three reasons: because it did best in the comparison just mentioned, because its questions are supposed to measure one single construct (the usability of a system) and because its questions are very subjective. The questionnaire has not been based on any set of existing guidelines, but it measures people's personal opinion about a system.

The SUS consists of 10 questions and yields one score ranging from 0 to 100.

The SUS used here had each word "system" replaced by "website", as was done in [Tul04].

The questionnaires used in this study contained the SUS for both websites (in arbitrary order). To make sure that each participant had a look at the two websites, two small questions about the contents were asked for each website.

Participants were also asked their age, gender, country, experience with looking for free software online and whether they had visited the two software websites before.

The two online questionnaires used exactly the same questions and differed only in format, so one questionnaire could be labeled as being professional and one as being unprofessional. For each participant, one of these two versions was randomly chosen. This choice was remembered for each computer, so returning visitors would get the same version.

The unprofessional questionnaire violated 10 of the 32 guidelines, while the professional one violated none. The most severe guideline violations of the unprofessional questionnaire were format inconsistency, a small font and the simultaneous use of infamiliar fonts. The precise guideline violations can be found in Attachment A.

Accessibility guidelines have not been violated too much. For example, one guideline states that the font size should be at least 10 points. The unprofessional questionnaire had a font size of 9, not 5. Such extreme violations would really make the questionnaire unaccessible and thus make it impossible to participate in the website assessment.

2.4 Participants
Visitors of a website about free software [Moo] were asked to participate in this survey. (Hence the subject of the two websites that were assessed.)

Participants were told that this survey was about the evaluation of two software websites. They were not told that there were two different questionnaires, because this knowledge might have influenced the results.

3. RESULTS
129 people completed the survey. The answers of nine people have been discarded because there were several indications that they had not seriously answered the questions. The most important indications for this were that the questions about the contents of the websites were not given and that the SUS scores were exactly 50, which is the result of giving all answers in the same column.

Of the 120 participants, 63 had the professional questionnaire and 57 had the unprofessional one. Also, 63 people (not exactly the same 63) were asked to rate the professional website first and 57 people were asked to rate the unprofessional website first.

In this chapter, four types of results will be discussed: characteristics of the participants, the ratings of the two websites, the effects of personal characteristics on these ratings, and of course the effect of the questionnaire.

For the statistical analyses, I've tested everything 2-tailed. I've used the common significance level of 5%, but I'm also giving the actual P-values in this paper. This way, people can see for themselves how significant the results are. Numbers have been rounded for reading convenience.

3.1 Participant characteristics

The majority of the participants was male (72%). The average age was 44.8 years (SD = 13.2), although the male participants were older (48.3, SD = 11.0) than the female participants (35.9, SD = 14.3). Figure 1 shows the age distributions of male and female participants.
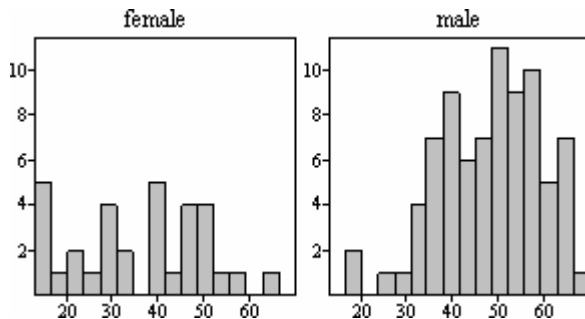

Figure 1. Age distributions of male and female participants

Almost half of the participants were from the USA (46.6%), while the UK, Canada and Australia delivered the most participants after the USA. In total, a large majority of the participants (78.0%) came from a country where English is one of the official languages. Table 1 shows the distribution of the countries and regions where participants came from.

Table 1. Country distribution of participants

Country / region	Number of participants	Percentage
USA	55	46.6%
UK	17	14.4%
Canada	7	5.9%
Australia and New-Zealand	7	5.9%
Europe (not UK)	15	12.7%
South- and Middle-America	6	5.0%
Asia	6	5.0%
Africa	5	4.2%
Not specified	2	

To the question how often participants looked for free software on the internet, a big majority answered "often" (45.4%) or "sometimes" (42.0%) as opposed to "seldom" (10.9%) or "never" (1.7%).

The majority of the participants (87.4%) had not visited one of the two websites before. Only 5.0% had visisted both websites before.

3.2 Ratings of the two websites

Figure 2 shows the distributions of the SUS scores for the two websites. The unprofessional website had a mean score of 59.6 (SD = 22.6), while the professional had 66.0 (SD = 21.7).
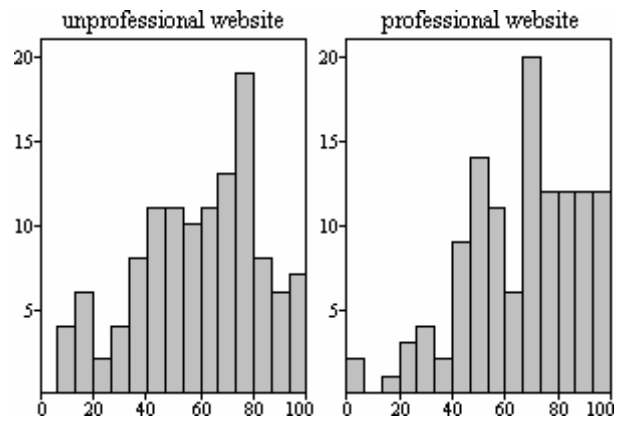

Figure 2. SUS score distributions of the websites

Most participants rated the professional website higher than the unprofessional one (56.7%), but a considerable group rated the unprofessional one higher (37.5%).

The difference between the website ratings in favor of the professional one (more precisely: the SUS score of the professional website minus the SUS score of the unprofessional website) had a mean of 6.4 (SD = 28.9). This difference is significant (t(119) = 2.4, P = 0.017).

3.3 Effects from personal characteristics

Age and gender did not correlate with the rating of either website.

Whether a website was visited before didn't significantly effect its SUS scores, although the P-value was 0.087 (t(117) = 1.7) for the professional website.

How often participants looked for free software on the internet, did not have an effect on the rating of the unprofessional website, but the professional website was rated significantly higher by more frequent software website visitors (F(4, 115) = 3.5, P = 0.009).

3.4 The effect of the questionnaire

For comparing the SUS scores between the two questionnaires, the total score for each participant will be used. The total score is the sum of the SUS scores for both websites.

Figure 3 shows the distributions of the total scores acquired with the different questionnaires. The unprofessional questionnaire yielded a mean total score of 121.2 (SD = 35.3), while the mean total score derived from the professional questionnaire was 129.4 (SD = 31.8).
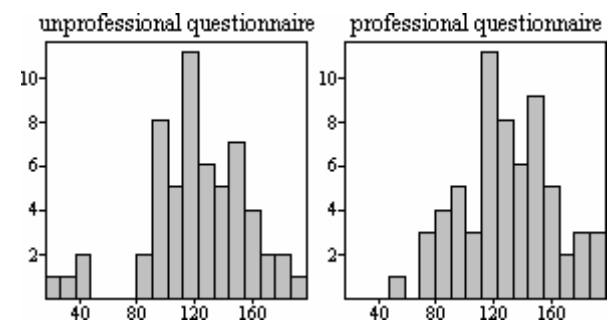

Figure 3. Total SUS score distributions of the questionnaires

The professional questionnaire yielded somewhat higher SUS scores, but this difference is not significant (t(118) = 1.3, P = 0.183).

For finding an interaction effect of the professionality of both the website and the questionnaire, all individual SUS scores are split into four groups, defined by the website rated and the questionnaire used. The mean scores of these four groups are shown in Figure 4. Although it's clear that both the website and the questionnaire made a difference in the scores, there seems to be no interaction effect.
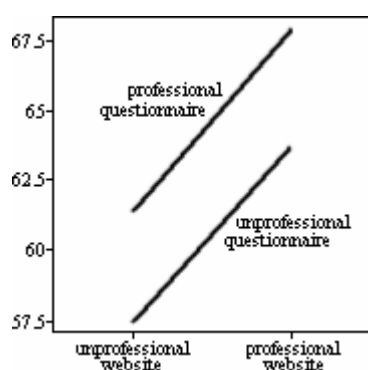

Figure 4. SUS score means

The order in which the two websites had to be rated, did not have a significant effect on the SUS scores. Although the first website was on average rated higher (65.1, SD = 20.7) than the second (60.5, SD = 23.8), this difference was not significant (t(119) = 1.7, P = 0.086).

4. CONCLUSIONS
4.1 Guidelines validity
A first conclusion that might be drawn, is that people seem to agree with common guidelines about website design.

The difference in ratings for both websites was significant, but here must be noted that the difference between these ratings was not very large. In fact, quite a lot of people (37.5%) rated the unprofessional website higher. A possible explanation is that the unprofessional website has some positive characteristics that the professional one doesn't have, like high accessibility of screenshots.

Nevertheless, the difference in website ratings was significant, indicating that people agree with the guidelines, at least about these two websites. The guidelines used in this study have passed another test of validity.

4.2 The effect of the questionnaire
The difference between the SUS scores obtained with the different questionnaires is not significant. These results suggest that the questionnaire format does not make a difference, or that the difference is very small.

If there is an effect of the questionnaire format, it's probably that people rate websites higher when they're using more professional questionnaires. The results of this study do not give convincing evidence for the existence of this effect, though.

4.3 Experience and navigation aids
Due to the small number of participants that had visited the websites before, it's hard to draw conclusions about their effects. The professional website was rated higher by returning visitors, although the P-value was just above the common 5%.

Perhaps this effect would have been significant if more returning visitors had participated.

This effect was far from significant for the unprofessional website. A plausible explanation is that the navigation aids of the professional website make it very accessible for users, especially when they return and know how to use them. The unprofessional website offers quite lousy navigation support, so returning visitors may have no advantage of knowing the website and its structure.

The fact that people who look for free software more frequently rate the professional website significantly higher, is another indication that the professional website offers something that experienced visitors appreciate, which might be the navigation aids.

4.4 The order in double website assessment
When two websites are rated by one participant, the results of this study suggest that people tend to rate the first website higher than the second one. The P-value of 0.086 is just above the common 5%, so there might be a slight tendency which can be found significant in a study with more participants.

5. DISCUSSION
5.1 This study in perspective
Many websites exist and most of them are quite different from each other. Also, the number of combinations of guidelines violations is virtually infinite, especially because each violation can be in different ways and to different degrees. For this reason, a lot of research with different websites and different guideline violation combinations is needed before we can draw firm conclusions about the effects of guideline violations, both with websites and with questionnaires, on website appreciation.

The current study is only one attempt to find certain effects, with only two websites, two questionnaires and one combination of guideline violations for these websites and questionnaires.

This study is an important step in this field and it should inspire future research to confirm or reject any conclusions or suggestions that this study has yielded.

Also, these effect might differ between countries and cultures. Since most participants of this study were English-speaking, middle-aged male, a generalization to other groups of people should be taken with care.

Whereas the conclusion about guidelines validity agrees with knowledge that has already been concluded from much other research, the conclusion about the effect of the questionnaire asks for more extensive research. Different studies with more participants should be able to find whether this effect is small or does not exist at all.

A considerable amount of speculation is involved in the conclusion about experience and navigation aids. The results might as well be explained by completely other factors. The given conclusion is just an explanation that seems very plausible, concerning the major differences between the two websites. This explanation may be tested by future research.

The conclusion about the order in double website assessment is based on insignificant results. This, too, should inspire future research to find out whether this effect exists or not.

5.2 Recommendations for future research
In the first place, as already mentioned, more research with different websites, questionnaires and participants is needed to gain more support for the conclusions of this study and to explore the scope to which the results can be generalized.

The conclusion about the order in double website assessment might also be an inspiration to research on assessment of three or more websites. If the order of two websites indeed influences the results of the assessment, the question arises how this influence would be for more than two websites.

5.3 Implications for questionnaire design

Since no effect of the questionnaire format has been found, questionnaire designers can design questionnaires in a way they like, without having to be afraid of any major effects that this will have on website assessment. Obviously, they are

recommended to appreciate the common guidelines, at least because this guarantees the accessibility of the questionnaire.

REFERENCES

[Aba04] Abascal, J., Arrue, M., Fajardo, I., Garay, N., & Tomás, J. (2004). *The use of guidelines to automatically verify Web accessibility*. UAIS, 2004, pp. 71-79.

[Abs] Absolutely Freeware: http://www.absolutelyfreeware.com

[Alp99] Alpar, P. (1999). *Satisfaction with a Web Site: Its Measurement, Factors and Correlates*. Electronic Business Engineering / 4. Internationale Tagung Wirtschaftsinformatik 1999.

[Bro96] Brooke, J. (1996). *SUS – A quick and dirty usability scale*. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor & Francis.

[Gle04] Gleitman, H., Fridlund, A.J., and Reisberg, D. (2004). *Psychology 6th edition*. W.W. Norton & Company, Inc., New York, New York, USA.

[Hac05] Hackett, S., Parmanto, B., & Zeng, X. (2005). *A retrospective look at website accessibility over time*. Behaviour & Information Technology, Vol. 24, No. 6, November-December 2005, pp. 407-417.

[Lum05] Lumsden, J. (2005). *Guidelines for the Design of Online-Questionnaires*. National Research Council Canada NRC/ERB-1127.

[Moo] Moor Computer Productions: http://www.moor-software.com

[Nie05] Nielsen, J. (2005). *Accessibility Is Not Enough (Jakob Nielsen's Alertbox)*: http://www.useit.com/alertbox/accessibility.html

[Onl] OnlyTheBestFreeware.com – Your source for the web's best freeware: http://www.onlythebestfreeware.com

[Pet04] Petrie, H., Hamilton, F., & King, N. (2004). *Tension, what tension? Website accessability and visual design*. W4A at WWW2004, May 18, 2004, New York, New York, USA.

[Res] Research-Based Web Design & Usability Guidelines: http://usability.gov/guidelines

[Sch62] Schachter, S., and Singer, J. (1962). *Cognitive, social and physiological determinants of emotional state*. Psychological review, 69, 379-399.

[Tul04] Tullis, T.S., & Stetson, J.N. (2004). *A Comparison of Questionnaires for Assessing Website Usability*. UPA 2004 Presentation.

[Van98] Vanderdonckt, J. (1998). *Towards a Corpus of Validated Web Design Guidelines*. Proceedings of the 4th ERCIM Workshop on 'User Interfaces for All', 1998, pp. 16-31.

[Zap05] Zaphiris, P., Ghiawadwala, M, & Mughal, S. (2005). *Age-centered Research-Based Web Design Guidelines*. CHI 2005, April 2-7, 2005, Portland, Oregon USA.

Attachment A – Guidelines used and violated

This table gives an overview of the 32 guidelines that have been used in this study, and whether these guidelines were respected by the websites and questionnaires that were used.

More extensive guideline descriptions have been omitted, but they can be found at [Res].

Guideline	Unprofessional website	Professional website	Unprofessional questionnaire	Professional questionnaire
Be consistent	No. Different pages have quite different design properties like font, alignment, navigation links and title positions.	Yes.	No. Color, font, and the location of the UT logo are not consistent between the different pages.	Yes.
Provide feedback to users	Yes.	Yes.	Yes. Although the total sequence of pages is not shown.	Yes.
Limit use of frames	Yes.	Yes.	Yes.	Yes.
Put important information at top of hierarchy	Yes.	Yes.	N/A.	N/A.
Use short sentence / paragraph lengths	Yes.	No. The descriptions of programs contain big paragraphs.	Yes.	Yes.
Provide page titles	No. Inconsistent and unsupportive page titles.	Yes.	No. All pages have the same title.	Yes.
Use well-designed headings	Yes.	Yes.	Yes.	Yes.
Determine scrolling vs. paging needs	Yes.	No. Some category pages are large, which	N/A.	N/A.

Guideline	Unprofessional website	Professional website	Unprofessional questionnaire	Professional questionnaire
		requires scrolling.		
Align page elements	Yes. Although not consistent.	Yes.	No. Pages are aligned, but not consistent.	Yes.
Reduce unused space	Yes. Although there is unused space for users with high resolutions.	Yes. Although there is unused space for users with high resolutions.	N/A. A questionnaire should be as clear as possible, without confusing information just to fill unused space.	N/A.
Put important information on top of page	Yes.	Yes.	Yes.	Yes.
Format for efficient viewing	Yes.	Yes.	Yes.	Yes.
Use readable font sizes	No. The font size is relative from the user's standard font size, but it is quite a lot smaller on most pages.	Yes. The default font size is 10 and it can be changed on the configuration page.	No. The font size is 9 points.	Yes.
Use familiar fonts	Yes.	Yes.	No. Serif and sans serif fonts are used simultaneously, in an inconsistent way.	Yes.
Enhance scanning	Yes. But, again, not consistent.	Yes.	Yes.	Yes.
Show links clearly	No. Links are normal text are not clearly	No. Not all the links look the same.	No. Links are in red and not	Yes.

Guideline	Unprofessional website	Professional website	Unprofessional questionnaire	Professional questionnaire
	distinct.		underlined.	
Indicate internal vs. external links	No. The links are very unclear and confusing.	Yes.	Yes.	Yes.
Use descriptive link labels	Yes.	Yes.	Yes.	Yes.
Use text links	No. Some links are images, with or without text.	No. But almost every link is a text link. The image links all contain text.	Yes.	Yes.
Repeat text links	No.	Yes.	N/A.	N/A.
Show used links	No. This is true for some links, but not all.	No. This is not true for any of the links.	No. This is not true for any of the links.	Yes.
Use graphics wisely	No. Some graphics have no clear purpose, like the monitor image at the top of the main page.	Yes.	No. The first page contains an image that has no clear use.	Yes.
Avoid graphics on search pages	N/A.	Yes.	N/A.	N/A.
Consider importance of search engine	Yes. There is no search engine, but this website doesn't have very many pages so a search engine is not needed.	No. This website has a search engine, which probably wouldn't be needed for this small number of pages.	Yes. There is no search engine, because it's not needed.	Yes. There is no search engine, because it's not needed.
Indicate search scope	N/A.	Yes. The scope of the search engine can be		

Guideline	Unprofessional website	Professional website	Unprofessional questionnaire	Professional questionnaire
		changed at the "Advanced search" page.		
Keep navigation aids consistent	No. Navigation aids are inconsistent and unsupportive.	Yes.	N/A.	N/A.
Group navigation elements	Yes.	Yes.	N/A.	N/A.
Place navigation on right	No. The navigation on the main page is in the middle and the category pages are lacking good navigation links.	No. The navigation elements are shown at the left and the top of each page.	N/A.	N/A.
Consider users' screen resolution	No. The category pages are slightly larger than 800x600.	Yes. The size of each page fits automatically to the browser size.	No. The size of each page is slightly too big for the 800x600 resolution.	Yes.
Use color wisely	Yes.	No. Some links might be a little hard to discover without good color vision.	Yes.	Yes.
Design for device independence	Yes.	Yes.	Yes.	Yes.
Provide alternative formats	No. Many images don't have alternative texts.	Yes. The images have alternative texts or the texts are given near the images.	No. The images have no alternative texts. But: there are only two images which are both not very important.	Yes. The one image on the questionnaire (the UT logo) has an alternative text.