Group versus Individual Web Accessibility Evaluations: Effects with Novice Evaluators

DOI:
10.1093/iwc/iww006

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Brajnik, G., Vigo, M., Yesilada, Y., & Harper, S. (2016). Group versus Individual Web Accessibility Evaluations: Effects with Novice Evaluators. *Interacting with Computers*, *28*(6), 843-861. https://doi.org/10.1093/iwc/iww006

Published in:
Interacting with Computers

doi:10.1093/iwcomp/xxxxxx

Group versus Individual Web Accessibility Evaluations: Effects with Novice Evaluators

Giorgio Brajnik[1] and Markel Vigo[2] and Yeliz Yesilada[3] and Simon Harper[2]

[1] *Dip. di Matematica e Informatica, Università di Udine, Udine, ITALY*
[2] *School of Computer Science, University of Manchester, Manchester, UK*
[3] *Middle East Technical University Northern Cyprus Campus, Güzelyurt, Mersin 10, TURKEY*

We present an experiment comparing performance of 20 novice evaluators of accessibility carrying out Web Content Accessibility Guidelines 2.0 conformance reviews working individually to performance obtained when they work in teams of two. They were asked to first carry out an individual assessment of a web page. Later on they were matched randomly to constitute a group of two and they were asked to revise their initial assessment and to produce a group assessment of the same page.

Results indicate that significant differences were found for sensitivity (inversely related to false negatives: +8%) and agreement (when measured in terms of the majority view: +10%). Members of groups exhibited strong agreement on the evaluation results among them and with the group outcome. Other measures of validity and reliability are not significantly affected by group-work.

Practical implications of these findings are that, for example, when it is important to reduce the false negatives rate then employing a group of two people is more useful than having individuals carrying out the assessment. Openings for future research include further explorations of whether similar results hold for groups larger than two, or what is the effect of mixing people with different accessibility background.

Categories and subject descriptors: teamwork; communication

Keywords: teamwork; user interfaces; web accessibility; accessibility evaluation

Responsible Editorial Board Member: xxx

1. RESEARCH HIGHLIGHTS

- When novice accessibility evaluators work in groups their ability to identify all the true problems increases (by 8%).
- Likewise, reliability of group evaluations increases (by 10%).
- Individual or group evaluations can be considered as equivalent methods with respect to false positives (if differences up to 8% in correctness are tolerated).
- Individual or group evaluations can be considered as equivalent methods with respect to overall effectiveness (if differences up to 11% in F-measure are tolerated).

2. INTRODUCTION

Web accessibility cannot be achieved only by following guidelines: there is a multitude of aspects that can affect the quality of accessibility evaluation results. Previous work has investigated several methodological aspects of web accessibility evaluations, ranging from sampling methods to evaluation techniques such as “barrier walkthrough” and conformance review, to the effect of expertise on the outcome of an evaluation (Brajnik et al., 2011, 2012), to definitions and perceptions of accessibility (Yesilada et al., 2014) and accessibility metrics (Vigo and Brajnik, 2011). All of these aspects can affect the outcome of an evaluation. Among them, the specific evaluation method being adopted plays an important role as well as the level of expertise of evaluators. For example, when considering the expertise effect on barrier walkthrough and WCAG 2.0 conformance review[1] it was found that expecting agreement at the 80% level, which was a level suggested for human testability, is not attainable when involving experienced or novice evaluators (Brajnik et al., 2011).

One factor that is expected to positively affect the results but that was not studied so far is the adoption of groupwise practices in evaluation: rather than having individuals performing independent evaluations of a web site, what happens when the same individuals are asked to interact while performing the assessment? As illustrated more deeply in the Related Work section, the following findings can be brought to bear on the question: (i) There are several factors that can potentially affect the performance of groups: if people interact, if people work on individual assessments or on group assessments, if people know each other, the size of the group, if people act as evaluators or as end-users, the expertise of people, if people work first individually and then in groups; (ii) Groups perform better than individuals on realistic settings; (iii) There are benefits of conducting groupwise usability evaluations over individual ones in terms of accuracy, correctness and sensitivity. (iv) These practices are being encouraged by standardization bodies and have been put in practice by the accessibility community. However, even if there are insights about the potential benefits of groupwise evaluation, nothing is known about its effectiveness when applied to accessibility.

In this paper we present results of an experiment involving 20 participants which was aimed at determining the effect of letting novice evaluators interact when using WCAG 2.0 to assess accessibility of given web pages. We decided to focus on novice evaluators and on small groups (of two people) because we argue that this is by far the most frequent situation: young inexperienced web developers that try to assess accessibility of what they develop; and should they ever work in groups, it is more likely that the group is small.

Our findings show that accuracy of results and the proportion of false positives are not significantly affected by working in groups. On the other hand, the ability to identify all true problems increases by 8% when working in groups; a corresponding increase in an overall measure of effectiveness (F-measure) of about 4% is achieved by groups. Max-agreement, one measure of reliability that represents the majority view, shows an increase of about 10% when people work in groups, while other reliability measures show no significant change. Analysis of how individuals changed their mind when working in groups shows that besides a large agreement (between participants of the group and the group outcome), the group effect leads both to a slightly higher proportion of false positives and of true positives.

In a previous research, when we adjoined two individual evaluations performed totally independently, the false negatives rate was 22%; this time, when two evaluators interacted in a group they achieved a rate of 35%. While it is difficult to draw practical implications (because false negatives can only be assessed after-the-fact and this is possible only if one assumes that the correct ratings are already known), the results obtained from the current research suggest that letting people interact decreases (by 13%) the *potential* ability to catch all the true problems. Thus, working in pairs improves such an ability as opposed to working alone, but the improvement is smaller than expected.

3. RELATED WORK

This section investigates the existing work that has been conducted to compare the performance of individuals when working alone and when within groups, especially with regards to usability and accessibility evaluation methods.

3.1. Individual vs Group Performance

A vast literature analyzes the problem solving performance of individuals and groups. The work by Hill shed some light on the confusion about individual *vs* group performance by unifying the terminology used and the tasks described by previous studies (Hill, 1982). Typically, in the analyzed studies participants have to apply their problem-solving skills to solve a task that requires a creative solution (*e.g.*, “The missionaries and cannibals” type problems (Jeffries et al., 1977)). Performance is usually measured in terms of task completion time, quality of solutions or number of trials for finding the solution. Hill’s review indicates that group performance is generally superior to the average individual performance, but inferior to the best individual in a group. Subsequent work (Miner, 1984) corroborated Hill’s review. Interestingly, this work revealed that an important factor is who the final judge is: in fact, when the best outcome was identified by the group, then the group outperformed individuals. However,

[1] A particular accessibility evaluation method based on using version 2.0 of a system of principles, guidelines and success criteria defined by the World Wide Web Consortium.

performance did not vary when the best individual decision is identified by an external observer (for example, the researcher). The order of the trials did not play any role: similar effects were found between group decisions that were not preceded by individual decisions as well as when they were preceded by individual decisions.

Later research challenged these results and found that groups always outperformed the best individual of the group (Michaelsen et al., 1989). But, for this to happen, the problem solving task, the group building criteria and the rewards needed to be more realistic than in previous works. For instance, participants have to know each other beforehand and they have to take their time to work on the problem. This is because extended periods of work boost trust and trustworthiness, increasing team performance. It was found that 40% of the groupwise gains cannot be explained by individual performance within the group, suggesting that groupwise work has a positive effect that exceeds the aggregated performance of individuals. Also, the prominence of the leading member of the group fades away as the task is prolonged over time (Watson et al., 1991).

3.2. Individual vs Group Performance on Usability Evaluation

In (Hornbæk and Frøkjær, 2008) the “evaluator effect” is discussed, *i.e.* the fact that usability evaluators in similar conditions identify substantially different sets of usability problems. It is argued that possible causes include vague evaluation procedures, variability of task scenarios, unclear problem descriptions, unclear criteria for what constitutes a usability problem, ambiguity regarding the matching procedure (such as who does the matching, those that carried out the evaluation or not? And is the matching performed individually or group-wise?).

Another source of confusion is the number of users (or evaluators) required to catch a meaningful number of usability problems – and possibly reduce the evaluator effect; this has been and still is a heated issue in the HCI community (Faulkner, 2003; Hwang and Salvendy, 2010; Schmettow, 2012). Nielsen (1992) showed that by pooling results from several evaluators the number of usability problems identified is increased. Similarly, Sears (1997) indicates that pooling results increases the evaluation's thoroughness compared to the results of a single evaluator. *Thoroughness* is the proportion of real problems found over all real problems, and it corresponds to our definition of *sensitivity*, which is explained in §4.3.

How group discussions, as opposed to individual evaluators, affect the output of usability inspection was also investigated by Følstad (2008). This study shows that 25% of the usability issues generated in the group discussions were new, *i.e.* they were not predicted by single evaluators. Furthermore, 25% of the issues found individually were eventually discarded and 37% of them were modified when discussed groupwise. This typically happened on those issues that were deemed of low severity. Similar results have also been reported by Hertzum et al. (2002): 11 usability experts inspected a web site individually and then discussed their findings groupwise. It was found that the actual overlap between the issues identified individually and groupwise was low (9% on average). However, despite of this low overlap, they perceived to be in agreement, which suggests that the groupwise method was more beneficial in increasing the confidence of evaluators than in removing the evaluator effect.

3.3. Individual vs Group Performance for Evaluating Web Accessibility

Accessibility of a web page can be assessed with different evaluation techniques (Abou-Zahra, 2008). These techniques can be broadly categorized into five classes: (i) inspection methods; (ii) automated testing; (iii) screening techniques; (iv) subjective assessments; (v) and user testing (Brajnik et al., 2012). When focusing on inspection methods, which are based on an evaluator inspecting a web page, there are many inspection techniques including heuristic evaluation, estimation, cognitive walkthrough, pluralistic walkthrough, feature inspection, consistency inspection, standards inspection or conformance evaluation and formal usability inspection (Nielsen, 1994). While some of these techniques are conducted by a single evaluator, some other may involve multiple evaluators. Feature inspection and standards inspection belong typically to the former group, whereas heuristic evaluation, cognitive walkthrough, pluralistic walkthrough and consistency inspection belong to the latter.

When it comes to web accessibility evaluation, the most widely adopted technique is conformance review or standards inspection (Yesilada et al., 2014). Conformance review is the technique by which the evaluator uses a set of accessibility guidelines or good practices that indicate how to prevent accessibility problems. The evaluator has to check whether a web page or web site meets these guidelines (Abou-Zahra, 2008; Thatcher et al., 2006; Henry and Grossnickle, 2007; DRC, 2004). As mentioned in the Introduction, when considering two methods, barrier walkthrough (Brajnik et al., 2011) and WCAG 2.0 conformance review (Brajnik et al., 2012), it was found that expecting agreement at the 80% level, which was a level suggested for human testability, is not attainable when involving either experienced or novice evaluators. Table 1 summarizes the scores obtained by novice evaluators for different validity metrics in these studies. “Cumulative” scores refers to joining the results that are produced by two *non interacting* evaluators (*i.e.*, considering the results of either of the two evaluators), hence where no group work is involved. It can be observed that cumulative scores are most of the times higher than those produced individually, meaning that the results of two independent evaluators are usually better than single ones.

Recently, Bailey et al. (2014) in an experiment on an accessibility evaluation method called "structured walkthrough", yielded confirmation for some of the figures shown in Table 1 concerning performance of individual novices: they found reliability (max-agreement) to be about 70% and accuracy to be about 64%.

Other studies challenge the effectiveness of WCAG conformance reviews. For example, Power et al. (2012) experimentally discovered that only about half of the problems encountered by blind users were covered by WCAG 2.0 success criteria.

However, in all these studies nobody investigated how the performance of evaluators is affected when they work in groups collaboratively, in particular when they get together to pool individual results. For WCAG conformance reviews, it has been suggested that collaborative approaches could give better and more reliable results (WAI, 2002, 2014). It is indicated that "it is less likely that one individual will have all the expertise that a collaborative approach can bring"; and further that "Using the combined expertise of different evaluators may sometimes be necessary or beneficial when one evaluator alone does not possess all of the required expertise." However, to this day, there is no empirical evidence that shows how the outcome is affected when a collaborative approach or a pooling technique to combine individual results is chosen compared to a single evaluator approach.

The idea of having multiple evaluators for assessing web accessibility is however not new. The rules of the AIR competition (Accessibility Internet Rally), which is organized by Knowbility since late '90s, prescribe that in a first round 2 evaluators independently review a web site and fill out a spreadsheet that computes a metric. In a second round, a more experienced judge corrects the initial assessments if the scores generated at the first stage differ more than 15 points. As a consequence of adjusting the initial results towards the "ground truth" – represented by the assessment made by quality control judge – one would expect more accurate ratings. Similarly, the procedure established by the German Government (BITV-Test[2]), introduced the idea of having two independent evaluators who grade on a 5 point scale how a given web page meets the 50 checkpoints comprised by the test. Then, the evaluators work together to reach an agreement on those criteria that got a different rating in the initial stage. It is claimed that this method increases accuracy and reliability, although no evidence supporting these claims is provided (Fischer and Wyatt, 2012).

3.4. Summary

Based on the analysis of this section we can conclude that: (i) There are several factors that can potentially affect the performance of individuals working in groups: if people interact, if people work on individual assessments or on group assessments, if people know each other, the size of the group, if people act as evaluators or as end-users, the expertise of people, if people work first individually and then in groups; (ii) Groups perform better than individuals on realistic settings (see section 3.1); (iii) There are benefits of conducting groupwise usability evaluations over individual ones in terms of accuracy, correctness and sensitivity (see section 3.2). (iv) These practices are being encouraged by standardization bodies and have been put in practice by the accessibility community. However, even if there are insights about the potential benefits of groupwise evaluation, nothing is known about its effectiveness when applied to accessibility (see section 3.3).

4. EXPERIMENTAL PLAN

To address the gap highlighted in the previous section, the main goal of our research is to determine the difference in performance of accessibility evaluators that use WCAG 2.0 conformance process when they work individually compared to when they work in small groups. In particular, in both cases participants carry out individual evaluations. But in one case they match their results in small groups and produce a group report, rather than an individual one. We focus on novice evaluators because it is known that their performance differs from that of more experienced evaluators (Brajnik et al., 2011, 2012) and they represent the most typical situation of developers inexperienced in accessibility trying to assess the accessibility shortcomings of what they develop. We also expect to see more impact when they work in groups.

We have seen in the Related Work section that, in general, accuracy produced by accessibility evaluators is relatively poor; a possible way to improve accuracy is to ask evaluators to work in groups so that joint exploration of pages and/or discussion of the issues that each individual evaluator finds could lead to better assessments. However, up to now, it is not clear if group-work has any effect on performance, how much it might affect performance, and if the effect depends on the

[2] Available at http://www.bitvtest.eu/bitv_test/intro/overview.html

Study	**Method**	**Accuracy**	**Correctness**	**Sensitivity**	**F-measure**	**Any-2 agreement**	**Max agreement**
(Brajnik et al., 2011)	BW	78	49	50	23–48	31–56	76–81
2 users, cumulative (Brajnik et al., 2011)	BW	–	–	71	–	–	
(Brajnik et al., 2012)	CR	66	56	55	52	54	72 (M)
2 users, cumulative (Brajnik et al., 2012)	CR	–	54	78	around 60	–	–

Table 1. Effectiveness scores (in percentage) obtained in previous studies (BW - Barrier Walkthrough, CR - Conformance Review). The entries "(2 users, cumulative)" mean that the results obtained by each of a pair of judges were joined without asking them to interact. (Definitions of mentioned concepts are given in § 4.3.)

experience of evaluators. In this research we decided to tackle only part of this question, and defer the study of whether evaluator experience has any effect to later investigations, to be carried out after some effect of group-work has been identified.

In section 3 we saw that there are many different accessibility evaluation methods (AEMs). In this experiment we considered two of them:

Individual evaluation (IE) A WCAG 2.0 conformance review, whereby a single evaluator considers a given set of WCAG 2.0 success criteria, and corresponding sufficient/advisory/failure techniques and decides if the pages being assessed conform or not to one of the three WCAG 2.0 conformance levels; see (W3C/WAI, 2008). This is also the same accessibility evaluation method that was adopted in previous experiments (such as (Fischer and Wyatt, 2012; Brajnik et al., 2012)).

Individual evaluations and merging (IEM) In this case evaluators follow a two-step procedure: they first individually assess conformance to WCAG 2.0 of pages, and later on they meet and jointly revise their assessments to reach consensus, by smoothing out differences and resolving conflicts. This is similar to what is done with usability heuristic evaluations (Nielsen, 2002). Notice that IEM subsumes IE, in that evaluators in the first step have to carry out an complete individual assessment.

For this experiment we used the same pages that were used in a previous study (Brajnik et al., 2012). This was to make it easier (for us and for subsequent researchers) to compare results obtained in this experiment with the previous one. The pages we considered are (i) "I love God Father movie" Facebook group; (ii) "The Godfather at IMDB"; (iii) "Bloomberg.com: WorldWide"; (iv) "Biotechnology News, Articles, and Information from Scientific American". We used stored versions of these pages on a Web server that was made available to participants. As previously, these pages were chosen because they differ in layout, complexity, genre and also in terms of accessibility support. Because we wanted to have a sufficiently large sample for each of these pages, we first assigned subjects to two pages, and then started assigning them to remaining pages; for this reason in the end participants did not evaluate the Facebook group page.

Because the goal of the experiment is to assess the effect of group-work, in this experiment we are not concerned about the effect that different pages might play, and therefore results will not be generalized against pages, but only with respect to participants (*i.e.*, the random factor is the evaluator, not the page being evaluated).

We considered level A and level AA WCAG 2.0 success criteria, 25 and 13 respectively (W3C/WAI, 2008)[3]. On one hand, these are the most important accessibility criteria and they are also the ones that are referred to by official regulations, such as the Italian accessibility law[4].

4.1. Participants

Participants were recruited from students attending a third-year undergraduate course on "User-Centered Web Development" taught by one of the authors in December 2012. Lectures included about 14 hours of web accessibility that followed about 40 hours of lectures on usability and user-centered development. All of them were invited to join the experiment, and were told that their outcome would be graded; after being told of the grades, they could decide whether their outcome would become part of the exam and concur to the final score; if they did not want that, then no negative effect on their exam would take place. No compensation was given to participants.

[3] WCAG 2.0 are organized in three levels of importance: level A include success criteria whose failure has potentially a high impact on users and which can be easily fixed; level AA and AAA include success criteria that either have a lesser impact or that are more difficult to satisfy.

[4] See `http://www.w3.org/WAI/Policy/`

4.2. Procedure

Students became aware of the experiment during lectures. They were invited to join the experiment, and those who did, were included in a list. They were then sent a demographic questionnaire, with 2 5-points Likert questions asking them about their knowledge in web accessibility and in WCAG 2.0, and additional questions regarding the number of sites that they had evaluated in the previous 6 months, if they did that using WCAG 2.0, and if they worked as accessibility consultants. In addition we asked them their age, gender, first language and if they had any disability.

They received also a spreadsheet containing the success criteria, grouped by guideline; the order of guidelines and of success criteria was randomized so that each participant received a different list. Each participant was also assigned to one of the test pages. For each success criterion (applied to the assigned page) the participant had to provide the outcome, in terms of `passed, failed, not-applicable`; he or she had also to rate the difficulty of such an assessment (5-point Likert scale), the WCAG 2.0 techniques that supported such a conclusion, and provide a brief explanation.

Finally, a post-hoc questionnaire had to be filled in which included questions asking how long the evaluation process took, which tools were used, if the page was already known, and four 5-point Likert scales asking to rate the effort needed, the perceived productivity, the ability to use WCAG 2.0, and the confidence in the outcome.

Participants were given 10 days to complete this first part of the work, which constitutes application of the IE method; they were asked to send via email all 3 questionnaires.

The second stage with the IEM method started after they did so. Participants were randomly grouped in pairs, provided that both members had evaluated the same page. They were given another copy of the spreadsheet with the success criteria (again guidelines and success criteria were randomized) and were asked to jointly re-evaluate the *same* page, by reconsidering what they individually found during the previous step, and by resolving possible conflicts. No particular method was suggested for conflict resolution. As before, they had to specify the outcome (`passed, failed, not-applicable`), difficulty, WCAG 2.0 techniques and possible explanations and comments. In the end they had to fill-in another copy of the post-hoc questionnaire that asked how long did this evaluation take, if there were conflicts and on which success criteria, and the four questions on effort, productivity, ability and confidence. Participants were given one week to complete this second step. According to the factors identified by Hill, what we did was asking people to interact while working on a group product (the joint evaluation), people didn't necessarily know each other before hand and they worked first as individuals and only then in groups of two; performance of the group was evaluated externally from the group, *i.e.*, by us (Hill, 1982).

4.3. Independent and Dependent Variables

This experiment has one independent variable which is the adopted accessibility evaluation method (two levels, *IE, IEM*). Other independent variables are the WCAG success criteria (38 levels, split into A and AA categories) and the page. Method and success criterion are within-subjects factors, whereas page is between-subject.

Usability evaluation methods have been studied thoroughly; but despite this, several problematic issues still remains about them. In their seminal work Gray and Salzman (1998) highlight several validity threats of research aimed at assessing properties about these methods. One of their suggestion is to try to adopt several converging measures to "triangulate" on the notion of usability (Gray and Salzman, 1998, pp. 242). For this reason we adopted several measures of effectiveness and of reliability of the two methods that we studied.

Dependent variables include *reliability*, which refers to the extent to which independent evaluations produce the same results (Brajnik et al., 2012, 2011). Reliability is important because if an evaluation method consistently leads to low reliability scores then applying it in reality is likely to produce results that cannot be reproduced, and that are affected by disturbance factors outside the control of the people applying the method. Reliability can be operationalized in several ways, grouped into two families: one that reflects variability (such as standard deviation of the scores or of the number of problems identified by individual evaluators); an example is reproducibility, used by Sears (1997). The other family is based on the agreement between evaluators, such as maximum agreement or any-two agreement. *Max-agreement* (MA) is defined as the relative frequency of the mode, *i.e.* the percentage of occurrence of the most frequent value of the set of ratings. Because the minimum value of MA is determined by the resolution scale of the ratings (for example, with ratings in {`pass`, `fail`, `not-applicable`} the minimum value for MA is 0.33, whereas for binary ratings the minimum value would be 0.5), one could also compute a linear adjustment to normalize MA within $[0, 1]$, so that 0 corresponds to the minimum value and 1 to 1. In this paper we report the unnormalized value of MA, which can be more easily interpreted.

Any-two-agreement is a further way to characterize the amount of agreement between judges; it focuses on the number of objects that were rated similarly by two random judges; this is a measure that is often used in reliability studies of usability evaluation methods, for example by Brajnik et al. (2012, 2011); Hertzum and Jacobsen (2001). Given a pair of judges that rated the same set of objects (in our case WCAG 2.0 success criteria), any-two-agreement ($A2$) is given by the ratio of the number of criteria on which the two ratings are the same over the total number of criteria. Given more than two judges, one can compute the mean value of $A2$ over all the possible unordered pairs of judges. We computed the mean $A2$ for all pairs of judges that evaluated a given page; therefore $A2$ is the mean of the proportion of criteria that were rated in the same way by any pair of judges.

Another dependent variable in the experiment is *accuracy*, which together with correctness, sensitivity and F-measure, is associated to the quality of the outcomes produced by participants; this is similar to what was done in a previous experiment (Brajnik et al., 2012), and indeed it rests upon some of its results. That experiment asked (among others) 25 experienced evaluators to apply WCAG 2.0 to the same pages we used here. In that experiment, a *correct rating* was a rating of a success criterion against a page such that the majority of experienced evaluators agreed on it; in other terms the values taken from {`fail`, `pass`, `not-applicable`} that, for each combination of page and success criterion, constitute the mode (the most frequent value). In case of ties all the modes were considered as correct.

For the current experiment, we adopted the same set of correct ratings of the previous experiment, relying in this way to the judgment provided by the majority of experienced evaluators.

Given a set of ratings, *accuracy* is the proportion of correct ratings. Because those depend of the kind of page and its features, other indexes of validity should be considered. After restricting to a given page, we can define the *true violations* (TV) for that page as the set of success criteria that are correctly rated as "fail"; the set of *found violations* (FV), given a participant and a page, is the set of ratings equal to "fail". These sets can be used to define three indexes:

Correctness $C = \frac{|TV \cap FV|}{|FV|}$ is the proportion of found success criteria violations that are also correct.

Sensitivity $S = \frac{|TV \cap FV|}{|TV|}$ is the proportion of all the true success criteria violations that were found. This matches the definition of thoroughness given by Sears (1997).

F-measure $F = \frac{2C \cdot S}{C+S}$ is the harmonic mean of C and S, a balanced combination of C and S summarizing validity of an evaluation. Neither correctness nor sensitivity alone can characterize validity, they have to be considered jointly: F-measure is a convenient way to provide an overall index of validity. Notice that a given change of x% in F-measure is equivalent to an x% change of correctness *and* a simultaneous x% change of sensitivity.

Notice that we addressed another recommendation that Gray and Salzman (1998, pp. 239) suggested, that is, to consider all possible outcomes of an evaluation: true-positives (which they call "hits"), true-negatives ("correct rejections"), false-positives ("false alarms") and false-negatives ("misses"). In fact we started with a golden notion of what a correct and an incorrect rating is, based on what a rather large set of experienced evaluators said, which allowed us to operationally fill-in all four of those categories.

4.4. Experimental Hypotheses

Based on the literature review, to support our overarching goal we formulated the following three hypotheses:

H1 Evaluators working in groups compared to individuals achieve higher accuracy results.

H2 Evaluators working in groups compared to individuals are more effective in terms of increased correctness and sensitivity (and, as a consequence, increased F-measure).

H3 Evaluators working in groups compared to individuals achieve more reliable results in terms of significantly different max-agreement and any two agreements.

H1 and H2 are inferred by the general results on group improvements of performance (§ 3.1) and on effects of group-based evaluations of usability (§ 3.2). H3 is inferred from general statistical considerations: if a population of n individuals is clustered in $k < n$ disjoint groups, then variability of the k groups outcome is expected to be smaller than that of the n individuals.

Practical implications of these hypotheses are important. Depending on the effect size of group-work, it may be worthwhile to team 2 or more novice evaluators to get better accessibility results. Furthermore, it could be worthwhile to team novices so that they learn from each other and more quickly get to speed and reach performance levels that are similar to those of experts.

5. RESULTS

About 40 students were invited to join the study; 20 accepted it (5 females) and completed the assigned tasks. Their mean age was 22.5 (SD=2.61), with a range from 21 to 33. For the second step, they were grouped into 10 groups. No student reported any disability.

Table 2 provides the distribution of the self-rated knowledge in accessibility and WCAG 2.0, for individuals and for corresponding groups (for a group we report the maximum value of the two members). The questions were

"Rate your knowledge in accessibility (1=very low, 5=very high)" and "Rate your knowledge in WCAG 2.0: (1=very low, 5: very high)". It can be seen that participants feel their knowledge in WCAG is weaker than that of accessibility, for which most of them state a neutral level (3). Also when looking at the same data groupwise, we see that the majority of groups feature a neutral level of knowledge.

It can be noted that our sample is quite homogeneous in terms of knowledge and experience.

The 20 participants, for the individual assessment (method=IE) produced a total of 924 ratings. Of these 19 ratings were incomplete because no success criterion outcome was given, and 49 lacked data about difficulty. We removed those 19 ratings from the dataset, leaving 459 for page `scientific`, 368 for `imdb` and 78 for `bloomberg`[5], for a total of 905 ratings.

Each of the WCAG success criteria had between 17 and 40 ratings, M=23.8. Each participant provided between 39 and 47 ratings (some success criterion was duplicated when the evaluator found 2 or more places where the success criterion could apply), M=45.2, SD=2.90. There were 297 `fail`, 300 `not-applicable` and 308 `pass`.

Tables A1 and A2 show the number of times each success criterion was rated as `pass`, `fail` or `not-applicable`. It can be seen that these numbers are relatively close to each other, indicating that each success criterion was being considered by participants. Notice that even some of the level A success criteria obtained relatively consistent scores (*e.g.*, 1.4.2, 3.2.1, 1.2.3), whereas other ones are associated to ambiguous outcomes (*e.g.*, 3.3.1, 1.2.1, 1.3.1). The same is true also for level AA success criteria: for example, 1.2.4 lead to unambiguous results, while 3.3.3, 2.4.7, 2.4.6 are among the most ambiguous ones.

Figures 1 and 2 show the mean difficulty levels (grouped per success criterion) that participants declared when assessing each success criterion (a 5-point Likert question, from -2 to 2). The figure shows also the grand mean for level A success criteria (M=-1.02) and for level AA (M=-0.95). Level A were found to be slightly easier to apply than level AA (but the difference is not significant). No participant stated that success criteria were more difficult than the neutral value.

As mentioned above, the 20 participants were grouped into 10 groups, which collectively provided 462 valid ratings. `Scientific` got 235 ratings, `imdb` 188 and `bloomberg` 39; Table A3 shows the distribution of ratings across success criteria and rating values.

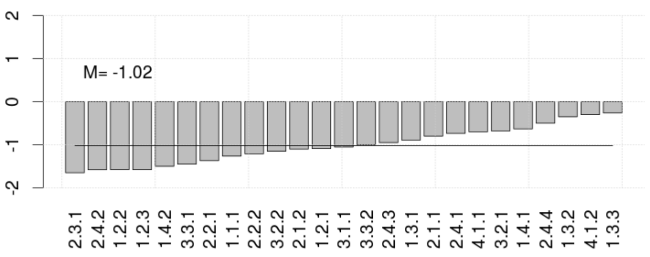

Figure 1. Difficulty levels for level A success criteria, in a scale from -2 to 2 (-2: Fully disagree, 2: Fully Agree - Statement: Evaluating the criterion was difficult).

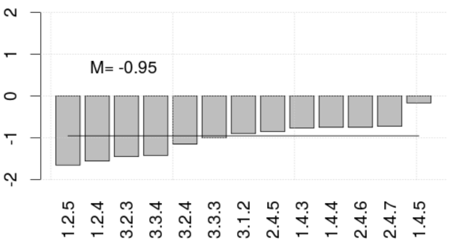

Figure 2. Difficulty levels for level AA success criteria, in a scale from -2 to 2 (-2: Fully disagree, 2: Fully Agree - Statement: Evaluating the criterion was difficult).

Rating	N	Tot	%	p-value	CI
Pass	220	308	71	< 0.0001	$[66.00, 76.34]$
Fail	155	297	52	0.49	–
NA	161	300	55	0.23	–

Table 3. Accuracy scores produced by individuals split by rating value (N: number of correct ratings; Tot: number of ratings; %: accuracy in percentage; p-value: p-value of the null hypothesis that accuracy is 50%; CI: 95% confidence interval for accuracy).

5.1. Accuracy

Individuals

Of the 905 valid ratings, individuals gave 536 correct ones, *i.e.* an accuracy of 59.2% (this proportion is significantly different than 0.5: $\chi^2(1) = 30.45, p < 0.0001$, 95% confidence interval $[0.56, 0.62]$, even though it shows a very poor performance).

When split by success criterion level, accuracy for level A is 350/566 (61.84%) and 186/339 for level AA (54.87%). The two proportions are barely significantly

[5] These differing numbers are due to the ways in which participants were assigned to pages; we wanted to make sure that at least 2 pages had enough data, and only when this criterion was fulfilled we started assigning people to additional pages.

Ratings:	**1**	**2**	**3**	**4**	**5**
Individuals					
Knowledge in accessibility	0	2 (10)	14 (74)	3 (16)	0
Knowledge in WCAG 2.0	4 (21)	5 (26)	9 (47)	1 (5)	0
Groups					
Knowledge in accessibility	0	0	7 (70)	3 (30)	0
Knowledge in WCAG 2.0	1 (10)	1 (10)	7 (70)	1 (10)	0

Table 2. Frequency (absolute and relative) of self-rated knowledge (1: lowest agreement, 5: highest agreement).

different ($\chi^2(1) = 3.98, p = 0.046$; CI of the difference: $[0.10, 13.85]$%).

When computed for individual evaluators, accuracy ranges from 42.55% to 71.79%, M=59.34%, SD=7.77%. The 25% and 75% deciles are Q1=55% and Q3=65%.

When computed for each success criterion, accuracy ranges from 5.00% to 100%, M=61.09%, SD=21.37%. Table A4 shows accuracy values for each success criterion.

Finally, when breaking down accuracy by possible rating values, we obtain the values shown in Table 3: accuracy is significantly different than 0.5 (*i.e.*, the case where evaluators could simply draw a coin to decide) only when rating is `pass`. Thus, deciding that `pass` is the outcome is generally a more accurate decision than opting for `fail` or `not-applicable`.

Groups

When working in groups, of the 462 valid ratings, evaluators gave 277 correct ones, *i.e.* an accuracy of 59.96% (this proportion is significantly different than 0.5: $\chi^2(1) = 17.92, p < 0.0001$, 95% confidence interval $[0.55, 0.64]$; Q1=0.54, Q3=0.65); a poor performance in this case too.

When breaking down accuracy by possible rating values, we obtain the values shown in Table 4: accuracy is significantly different than 0.5 (*i.e.*, the case where evaluators could simply draw a coin to decide) only when rating is `pass`. Thus, also for groups deciding that `pass` is the outcome is generally a more accurate decision than opting for `fail` or `not-applicable`.

When split by success criterion level, accuracy for level A is 183/286 (63.99%) and 94/176 for level AA (53.41%). The two proportions are significantly different ($\chi^2(1) = 4.65, p = 0.031$; CI of the difference: $[0.88, 20.27]$%).

Rating	**N**	**Tot**	**%**	**p-value**	**CI**
Pass	102	138	74	< 0.0001	$[65.63, 80, 84]$
Fail	90	171	53	0.54	–
NA	85	153	56	0.20	–

Table 4. Accuracy scores produced by groups split by rating value (N: number of correct ratings; Tot: number of ratings; %: accuracy in percentage; p-value: p-value of the null hypothesis that accuracy is 50%; CI: 95% confidence interval for accuracy).

When computed for each group, accuracy ranges from 51.06% to 74.36%, M=60.20%, SD=7.28%.

When computed for each success criterion, accuracy ranges from 0% to 100%, M=61.80%, SD=25.47%. Table A5 shows accuracy values produced by groups for each success criterion.

5.1.1. *Individual* vs *groups*

As apparent from reported values, the difference in accuracy when working alone (59.23%) and working in groups (59.96%) is negligible and not significant. When considering only level A success criteria, we found no significant difference in accuracy due to individual/group either; similarly for level AA. Thus adopting method IEM instead of IM leads to no increase in accuracy.

Table 5 shows the set of 25% success criteria that achieved the highest accuracy for individuals and/or for groups, split by those which appear in both cases and those not. Similarly, Table 6 shows the worst 25% success criteria according to accuracy.

Some of the comments provided by judges suggest why these success criteria have so low scores. For example, when applying 1.4.4 judges gave contrasting opinions on whether to consider a page as responsive or not (depending on the degree of responsiveness), and on what negative consequences this could bring to people with disabilities. Furthermore, judges used different browsers that had different capabilities regarding text zooming. For 2.2.2 judges were unsure about whether the success criterion applies also to banner ads, of if they should be excluded from the assessment. For 2.4.7 questions arose because different browsers showed different default behavior with respect to how focussed elements in the page are highlighted. For 3.2.3 some judges were strict and marked it as `not-applicable` (because the analysis was limited to a single page); others explored nearby pages and discovered that some menu options were not consistently ordered. For 3.3.1 some judges overlooked a small form for entering an email address. For 1.4.5, some judges commented that because some of the CSS code was embedded in the HTML, the success criterion failed;

	Success Criterion	Level
	Topmost in individual and group evaluations	
1.1.1	Non-text Content: All non-text content that is ...	A
1.2.4	Captions (Live): Captions are provided for all ...	AA
2.4.2	Page Titled: Web pages have titles that describe ...	A
3.3.4	Error Prevention (Legal, Financial, Data): For Web ...	AA
4.1.2	Name, Role, Value: For all user interface components ...	A
	Topmost in individual but not in group evaluations	
1.4.3	Contrast (Minimum): The visual presentation of text ...	AA
2.1.1	Keyboard: All functionality of the content is ...	A
2.4.3	Focus Order: If a Web page can be navigated sequentially ...	A
2.4.5	Multiple Ways: More than one way is available to ...	AA
4.1.1	Parsing: In content implemented using markup language ...	A
	Topmost in group but not in individual evaluations	
3.1.1	Language of Page: The default human language of ...	A

Table 5. Topmost success criteria according to accuracy for the two evaluation methods.

others focussed on some image that contained text (such as movie icons in `imdb`) then the success criterion was to be rated with a fail. For this success criterion some judges were even more confused regarding sufficient and necessary techniques.

We argue therefore that the reasons why success criteria got low accuracy scores are due to potential difficulty in interpreting a success criterion, to ambiguity due to the way in which the analysis was carried out, and ambiguity in deciding to what parts of a page they should be applied to. Part of this might be due to the intertwined nature of usability and accessibility issues: for example, web sites not providing sufficient feedback (Power et al., 2012). However, the likelihood that this has happened with our participants is low, because students were not instructed to bring usability issues into the analysis, and guidelines themselves do not promote such insights. And furthermore, no qualitative comment provided by students referred to usability.

An ANOVA for testing if method, correctness of ratings, and level of success criterion have some effect on difficulty (as rated by evaluators, individually and groupwise) shows that the only significant main effect is due to method ($F(1) = 16.200, p < 0.0001$). Although significant, this difference (M=-1.00 for individuals and M=-1.22 for groups, in a scale $[-2, 2]$) can be easily explained by noting that by design groupwise evaluations always followed individual ones, and therefore it is to be expected that participants rate the first evaluation as more difficult than the second one. No significant interactions were found.

5.2. Correctness

Correctness is the proportion of success criteria violations reported by an evaluator that are correct. To compute correctness scores, we grouped all the ratings by evaluator (and consequently by page too). For each of these subsets of data, we computed the success criteria that were reported as violated (all the instances of a success criterion whose outcome was `fail`), and the fraction of these that were correctly labeled as `fail`. Computing the ratio between the size of these two sets gives the correctness score for that evaluator.

Across all 20 participants, correctness ranges from 0.29 to 1.00, M=0.56, SD=0.15, Q1=0.50, Q3=0.63. It is normally distributed (Shapiro-Wilk's normality test, $W = 0.922, p = 0.11$, suggests to accept the null normality hypothesis), with a 95% CI=$[0.49, 0.64]$.

For the ratings produced by groups we computed correctness in the same way. In this case it ranges from 0.40 to 0.73, M=0.54, SD=0.09, Q1=0.51, Q3=0.57. Again normality is assumed (as per Shapiro-Wilk's test), leading to CI=$[0.48, 0.61]$.

To compare correctness of IE *vs* IEM evaluations, we paired the data so that correctness of a group and correctness of the average of its two members could be compared. An ANOVA (treating evaluation method as a within-subjects factor) shows that no significant effect is caused by method, as could be expected from seeing how close the two means are (Bartlett's test was used to test homogeneity of variance).

To test equivalence we used the two one-sided t-tests method (equivalence testing means to determine how large a difference is deemed acceptable for the two groups to be considered the same). The null hypothesis is that the two means are *different*, in this case mean correctness for IE and for IEM. Application of such a test to the difference between correctness obtained by individuals and by corresponding groups leads to a significant result ($p = 0.049$), provided that the threshold for considering

	Success Criterion	**Level**
	Bottom-most in individual and group evaluations	
1.4.4	Resize text: Except for captions and images of text ...	AA
2.2.2	Pause, Stop, Hide: For moving, blinking, scrolling ...	A
2.4.7	Focus Visible: Any keyboard operable user interface ...	AA
3.2.3	Consistent Navigation: Navigational mechanisms that ...	AA
3.3.1	Error Identification: If an input error is automatically ...	A
	Bottom-most in individual but not in group evaluations	
3.1.2	Language of Parts: The human language of each pass ...	AA
	Bottom-most in group but not in individual evaluations	
1.3.2	Meaningful Sequence: When the sequence in which ...	A
1.4.5	Images of Text: If the technologies being used can ...	AA

Table 6. Bottom-most WCAG 2.0 success criteria according to accuracy for the two evaluation methods.

equivalent the two means is at least 0.081. In other words, when differences in correctness up to 0.081 (*i.e.*, 8%) are considered negligible, then the two methods can be safely considered equivalent with respect to correctness. Figure 3 shows the distribution of the scores.

To test whether the success criteria level played any role, we split the data into four subsets, one for each combination of the two levels (A and AA) and the two methods (IE and IEM). For each of these subsets we computed correctness (for each individual evaluator and for each group). On the resulting data set (on which Bartlett's test supported rejection of the hypothesis that variance is not homogeneous: $K^2(1) = 0.815, p = 0.367$) a repeated measures ANOVA was carried out to determine whether level and method have any effect on correctness. Only a main effect of level was found ($F(1) = 23.76, p < 0.0002$); no interaction with method and no effect of method can be reported. For level A correctness is M=0.63, while for level AA it is M=0.45; the difference is significant ($t(39) = 5.175, p < 0.0001$), with a CI $=[0.11, 0.25]$.

5.3. Sensitivity

Sensitivity, which is the proportion of correct success criteria violations that each participant reported, was computed in the same way as correctness.

Across all 20 participants, sensitivity ranges from 0.19 to 0.77, M=0.57, SD=0.20, Q1=0.47, Q3=0.74. It is not normally distributed (Shapiro-Wilk's normality test, $W = 0.853, p = 0.0061$, suggests to reject the null normality hypothesis). A bootstrap method (to cope with non-normality) leads to a 95% CI=$[0.49, 0.66]$.

For the ratings produced by groups we computed sensitivity in the same way. In this case it ranges from 0.12 to 0.92, M=0.65, SD=0.22, Q1=0.62, Q3=0.75. Again normality cannot be assumed (as per Shapiro-Wilk's test); the bootstrap method leads to CI=$[0.52, 0.77]$.

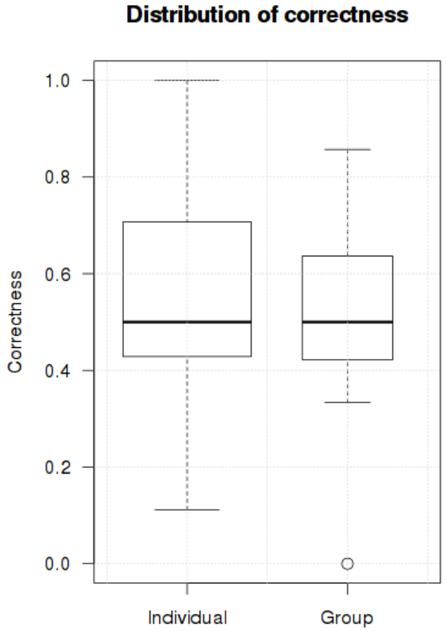

Figure 3. Boxplot of correctness obtained by individual evaluators and by groups.

As before, to compare sensitivity of IE *vs* IEM evaluations, we paired the data so that sensitivity of a group and sensitivity of its two members could be compared. An ANOVA (treating evaluation method as a within-subjects factor) shows that a significant main effect is played by method: $F(1, 19) = 5.32, p = 0.0325$ (Bartlett's test was used to test homogeneity of variance: $K^2(1) = 0.0372, p = 0.85$). Thus, the individual mean M=0.57 and the group mean M=0.65, when considering individual variability (*i.e.*, when pairing data), are significantly different; see also Figure 4

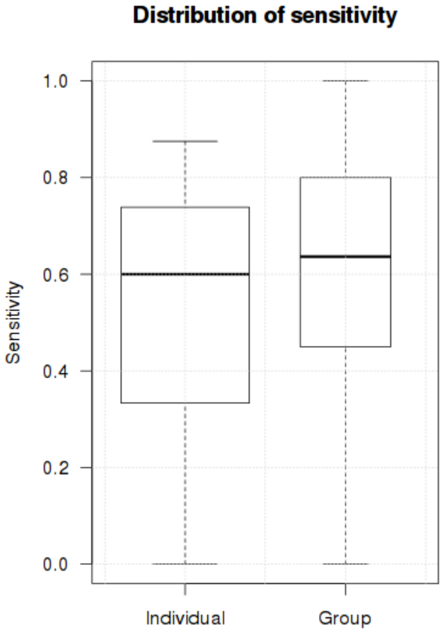

Figure 4. Boxplot of sensitivity obtained by individual evaluators and by groups.

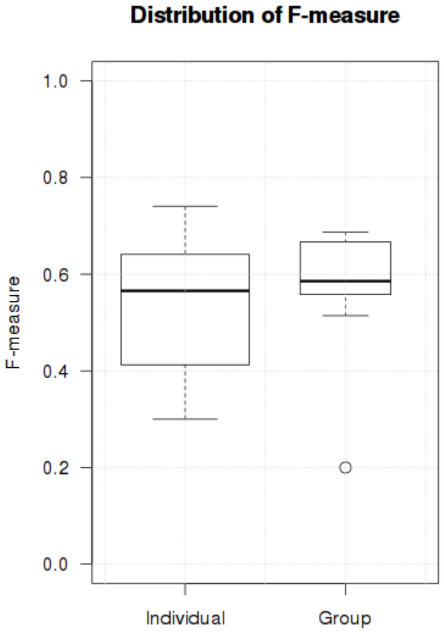

Figure 5. Boxplot of F-measure obtained by individual evaluators and by groups.

Also in this case we split the data into four subsets, one for each combination of the two levels (A and AA) and the two methods (IE and IEM). For each of these subsets we computed sensitivity. On the resulting data set (on which Bartlett's test supported rejection of the hypothesis that variance is not homogeneous: $K^2(1) = 0.0198, p = 0.888$) a repeated measures ANOVA was carried out to determine whether level and method have any effect on correctness. Besides a main effect of level ($F(1,19) = 55.79, p < 0.0001$), a main effect of method was found ($F(1,19) = 5.497, p = 0.0301$); no interaction of level with method can be reported. For level A sensitivity is M=0.68, while for level AA it is M=0.48, with a difference CI $=[0.10, 0.30]$.

5.4. F-measure

F-measure is the harmonic mean of correctness and sensitivity. Across all 20 participants, F-measure ranges from 0.30 to 0.74, M=0.53, SD=0.13, Q1=0.43, Q3=0.63. It is normally distributed (Shapiro-Wilk's normality test, $W = 0.946, p = 0.3081$, suggests to accept the null normality hypothesis), with a 95% CI=$[0.47, 0.60]$.

For the ratings produced by groups, F-measure ranges from 0.20 to 0.69, M=0.57, SD=0.14, Q1=0.56, Q3=0.66. It is not normally distributed (as per Shapiro-Wilk's test: $W = 0.738, p = 0.0025$); application of the bootstrap method leads to CI=$[0.48, 0.65]$.

As before, to compare F-measure of IE *vs* IEM evaluations, we paired the data so that F-measure of a group and F-measure of its two members could be compared. An ANOVA (treating evaluation method as a within-subjects factor) again shows that no significant effect is caused by method (Bartlett's test was used to test homogeneity of variance).

Application of equivalence test to the difference of F-measure obtained by individuals and that obtained by corresponding groups leads to a significant result ($p = 0.048$), provided that the threshold for considering equivalent the two means is 0.11. In other words, when differences in F-measure up to 0.11 are considered negligible, then the two methods can safely be considered equivalent with respect to F-measure. Figure 5 shows the distribution of the scores.

And finally, as before, we computed F-measure for each combination of success criterion level and method. On the resulting data set (on which Bartlett's test supported rejection of the hypothesis that variance is not homogeneous: $K^2(1) = 0.1203, p = 0.7287$) a repeated measures ANOVA was carried out to determine whether level and method have any effect on F-measure. Only a main effect of level was found ($F(1) = 84.39, p < 0.0001$);

no interaction with method and no effect of method can be reported. For level A F-measure is M=0.62, while for level AA it is M=0.50; the difference is significant ($t(39) = 10.45, p < 0.0001$), with a CI =[0.17, 0.25].

5.5. Reliability - Max agreement

Here we discuss reliability when measured as max-agreement, that is the unnormalized relative frequency of the mode of the ratings.

For individual data, max-agreement over the 38 success criteria ranges from 0.50 to 1.00, M=0.79, SD=0.20. Shapiro's test indicates non-normality of the data, the bootstrap method produces a 95% CI=[0.75, 0.83].

For ratings produced by groups, max-agreement ranges from 0.50 to 1.00, M=0.89, SD=0.16. Also here Shapiro's test indicates non-normality of the data, the bootstrap method produces a very high 95% CI=[0.86, 0.92].

To compare the two sets of values, across the two methods, we used the non parametric Wilcoxon rank test, which indicates a significant difference ($W = 4604, p < 0.0001$) and confirms the disjoint confidence intervals found with the bootstrap method. Application of such a method to compute confidence intervals for the difference between max-agreement due to the two methods yields CI=[0.05, 0.15]. Figure 6 shows the distribution of the two sets of values.

When splitting the data by success criteria level, for level A max-agreement for individual evaluators ranges from 0.50 to 1.00, M=0.80, SD=0.20, CI=[0.75, 0.83]. For the same level, max-agreement for groups ranges from 0.50 to 1.00, M=0.89, SD=0.15, CI=[0.86, 0.92]. The difference is significant (Wilcoxon's $W = 2080, p = 0.0029$), CI=[0.04, 0.15].

For level AA max-agreement for individual evaluators ranges from 0.50 to 1.00, M=0.77, SD=0.20, CI=[0.75, 0.83]. For the same level, max-agreement for groups ranges from 0.50 to 1.00, M=0.89, SD=0.17, CI=[0.86, 0.92]. The difference is significant (Wilcoxon's $W = 496, p = 0.0050$), CI=[0.04, 0.20].

5.6. Reliability - Any-two agreement

Any-two agreement is the fraction of success criteria that any two evaluators assessed and agreed upon. To compute it, we split the data into three sets, one per each page. For each subset we created all the possible different unordered pairs of evaluators, and for each pair we computed the list of success criteria where they agreed on the rating, and computed the ratio between this number and the total number of success criteria that both evaluators assessed. In case of multiple ratings for the same success criteria we adopted a conservative stance, and checked whether at least one was the same. At the end, we computed the average score over all pairs.

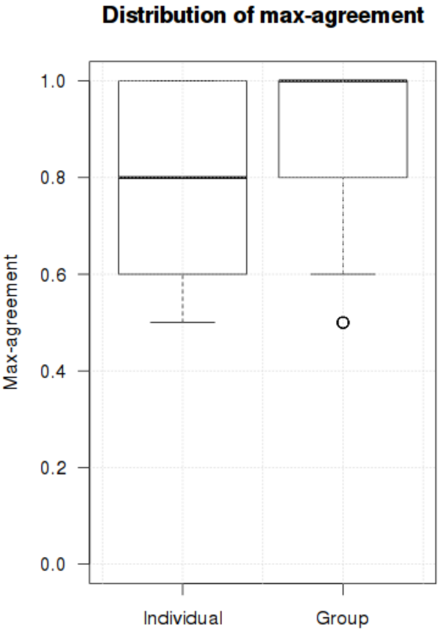

Figure 6. Boxplot of max-agreement obtained by individual evaluators and by groups.

When doing so with individual data we found 1 pair for `bloomberg`, 28 for `imdb` and 45 for `scientific`. Any-two agreement ranges from 0.34 to 0.95, M=0.67, SD=0.15, CI=[0.64, 0.71].

When computing any-two agreement for groups, after removing data regarding `bloomberg`, we found 6 pairs for `imdb` and 10 for `scientific`. Any-two agreement ranges from 0.58 to 0.87, M=0.72, SD=0.09, CI=[0.67, 0.77].

Both data are normally distributed (Shapiro's test: $W = 0.97, p = 0.077$, and $W = 0.93, p = 0.289$) and a t-test for comparing their means gives only a marginal difference ($t(33.71) = 1.7924, p = 0.082$).

Conversely, the test for equivalence supports rejection of the null hypothesis that they differ (assuming a difference up to 0.10: $p = 0.0481$). Hence, we can safely assume that there is no difference due to method. Figure 7 shows the distribution of any-two agreement.

When restricting only to level A success criteria, for individual data we get any-two agreement ranging from 0.42 to 0.92, M=0.68, SD=0.15, CI=[0.64, 0.72], whereas any-two agreement for groups ranges from 0.56 to 0.92, M=0.73, SD=0.12, CI=[0.67, 0.79].

For level AA and individuals, any-two agreement ranges from 0.19 to 1.00, M=0.65, SD=0.19, CI=[0.60, 0.69],

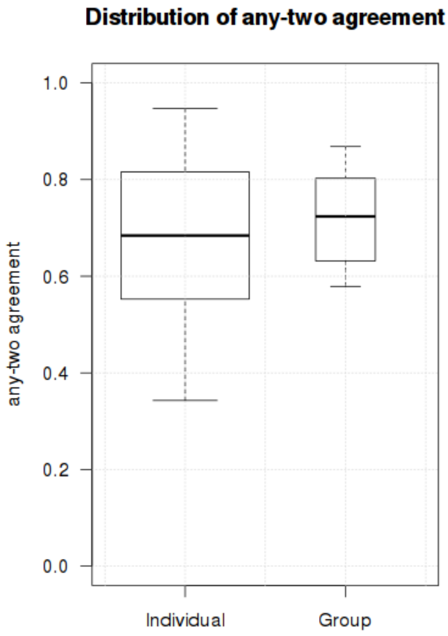

Figure 7. Boxplot of any-two agreement obtained by individual evaluators and by groups.

whereas for groups it ranges from 0.54 to 0.92, M=0.71, SD=0.12, CI=$[0.65, 0.77]$.

Comparing (through a t-test) the means of any-two agreement within each of the success criteria levels and between the two methods, yields a non-significant difference for level A, and a marginally significant difference for level AA ($t = 1.7085, p = 0.0965$). There are no sufficient data points to test for equivalence, for either of the success criteria levels.

5.7. Intragroup analysis

With the collected data we can analyze if there is any association between what was reported by individuals compared to what was reported by the group they constituted. In particular we classified the outcome of each rating as being a *true-positive*, a *false-positive*, a *true-negative* or a *false-negative* and then looked at how the two members of a group reached a collective outcome. In reading Table 7 consider that entries "tn" and "tp" correspond to favorable outcomes (where groups skipped non-existing problems and found existing ones); and vice versa, "fn" and "fp" correspond to unfavorable cases: missed and non-existing problems.

Because several of the 905 valid individual ratings involved repeated applications of the same success criterion on the same page by the same judge, and because we have no ways to compare these ratings across different individuals or between an individual and the group (*i.e.*, if evaluator P1 found 3 images failing success criteria 1.1.1, and evaluator P2 found on the same page 2 images that failed that success criterion, then we are unable to know which one of the first three images correspond to which one of the second set), then we decided to filter out all the duplicated ratings that within a page and judge refer to the the same success criteria. This left 612 individual ratings. For group ratings, from 462 valid ratings, filtering duplicates left 306 ones.

Reported	Expected	Outcome	
fail	fail	true-positive	tp +
fail	notapp	false-positive	fp -
fail	pass	false-positive	fp -
pass	fail	false-negative	fn -
pass	notapp	true-negative	tn +
pass	pass	true-negative	tn +
notapp	fail	false-negative	fn -
notapp	notapp	true-negative	tn +
notapp	pass	true-negative	tn +

Table 7. Decision table for outcomes; {tp, tn} represent correct decisions.

ind/grp	fn	fp	tn	tp	Total
fn	**19**	-	-	23+	42
fp	-	**73**	20+	-	93
tn	-	29-	**328**	-	357
tp	7-	-	-	**113**	120
Total	26	102	348	136	**612**

Table 8. Distribution and cross-tabulation of individual and group outcomes (pluses and minuses indicate a fortunate/unfortunate change due to working in group; dashes indicate an impossible combination).

To see if there is any pattern between the outcomes of the two members of the group and the corresponding group outcomes, we paired the outcomes of the two members and split the data according to all the unordered pairs of the possible outcomes.

Table 8 shows the cross tabulation of the 612 individual ratings against the 306 ratings produced by groups. Notice that some entries in the table are empty because of consistency: for example, it is impossible that, given a success criterion and page, ratings by judges or groups can be simultaneously fp and fn. Consistent combinations are only {fp, tn} and {fn, tp}. As shown by Table 8, of all the 612 individual ratings, the most frequent outcome is tn (with 357 occurrences), followed by tp (120), fp (93) and fn (42). For groups we get the same ranking: the most frequent outcome is tn (348), tp (136), fp (102)

judge-1	judge-2	group	number	change
fp	fp	tn	1	✓
fn	fn	tp	3	✓
fn	fn	fn	6	=
fn	tp	fn	7	✗
fn	tp	tp	17	✓
fp	tn	tn	18	✓
fp	fp	fp	22	=
fp	tn	fp	29	✗
tp	tp	tp	48	=
tn	tn	tn	155	=
	Total		306	

Table 9. Frequency of the group outcome depending on the outcome of the individuals.

and fn (26). Diagonal entries represent the frequency of agreement between either member of the group and the group, and show a high level of agreement: ($\chi^2(9) = 1031, p < 0.0001, Cramer's\ \phi = 0.75$). This means that working in group did not foster a change in a large number of ratings (the diagonal values in the table sum up to 533, *i.e.* 87% of all the ratings). Remaining entries show the number of times that an individual changed his/her mind when producing the group output (79, or 13%). Of these some turned out to be unfortunate decisions, leading to wrong answers (36, or 6%), and the remaining ones (43, or 7%) to correct decisions. Thus, there appears to be a small prevalence of correct over incorrect decisions taken as an effect of interacting; however this difference is not statistically significant.

Table 9 shows how the pairs of outcomes produced individually correspond to the group outcome. It can be seen that in 29 cases over a total of 306 group ratings (*i.e.*, 9%) when one evaluator marked a false positive and the other a true negative, the former prevailed. Conversely, in 18 cases (6%) when one thought of a false positive and the other of a true negative, the latter prevailed. In 17 cases (6%) a false negative and a true positive turned out to be a true positive. Thus it appears that overall the number of times that working in group improved the results (39) is similar to when it led to wrong decisions. When both participants made a wrong decisions (*e.g.*, fp) the group outcome was almost always fp.

6. DISCUSSION

As we have discussed in the Related Work section (§ 3), even though the existing work shows the benefits of groupwise evaluation for usability and the accessibility community recommends it, there is no research that investigates its effectiveness. Our work aims to fill that gap and in particular the conducted experiment investigated how the effectiveness and reliability affected by the groupwise evaluation. Here we discuss our findings in terms of the hypotheses we have stated in § 4.4.

6.1. Accuracy

Hypothesis **H1** states that evaluators working in groups compared to individuals achieve better results, in terms of increased accuracy. Data, however, do not support it. Accuracy is not influenced by the evaluation method: 59.23% and 59.96% of accuracy is obtained individually and groupwise respectively, which is not a significant difference. Considering that evaluators were supposed to resolve the evaluation conflicts they had, this may be an indication of (1) evaluators had almost similar ratings and therefore did not have to discuss anything so evaluations were not changed; (2) there was a counterbalancing effect when modifying their original evaluations: some wrong ratings were corrected and some correct ratings were turned into wrong ones. Indeed, data from § 5.7 suggest that in the vast majority of cases individuals did not change much (87%), which is compatible with the relatively high figures we got for reliability (max-agreement being 0.79 and 0.89, any-two agreement being 0.67 and 0.72). The counterbalancing effect also occurred, but in much smaller number of cases: +46 *vs* -36. We propend therefore for the former explanation, also after considering that evaluators had the same background and experience, and thus could not contribute varied experiences to the group.

If we look deeper and compare accuracy at A and AA levels, there is an increase of 2 percentage points groupwise for A level success criteria (62% for individual evaluations *vs* 64%), whereas there is a decrease of another 2 percentage points for AA (55% *vs* 53%); both differences are not statistically significant though. This suggests that any expected complexity in success criteria does not interact with adoption of a group method.

Regarding the counterbalancing effect, Table 10 contains the top 5 success criteria that obtained the highest levels of accuracy for individual evaluations, while Table 11 contains those that got the lowest scores for accuracy.

These tables clarify that the counterbalancing effect might also happen across success criteria, as follows: those success criteria that have higher levels of accuracy at individual level feature also a high groupwise accuracy, whereas for success criteria that have a low accuracy individually, groupwise evaluation tend to decrease their accuracy. Therefore groupwise evaluations produce the same overall accuracy although the worst and best SC tend to get closer to both ends.

SC	description	level	individually	groupwise	diff
2.4.2	Page titled	A	1	1	0
1.1.1	Non-text content	A	1	1	0
3.3.4	Error prevention	AA	0.95	0.9	-0.05
1.2.4	Captions (live)	AA	0.95	1	+0.05
4.1.2	Name, role, value	A	0.85	1	+0.15

Table 10. Most accurately rated success criteria (SC) and the difference between individual and groupwise accuracy.

SC	description	level	individually	groupwise	diff
2.2.2	Timing adjustable	A	0.05	0	-0.05
3.3.1	Error identification	A	0.10	0.10	0
3.2.3	Consistent navigation	AA	0.25	0	-0.25
2.4.7	Focus visible	AA	0.35	0.25	-0.10
3.1.2	Language of parts	AA	0.42	0.50	+0.08

Table 11. Least accurately rated success criteria.

6.2. Effectiveness

Hypothesis **H2** states that evaluators working in groups compared to individuals are more effective in terms of increased correctness and sensitivity (and, as a consequence, increased F-measure). Our data partially supports this – we see an increase in sensitivity but not in correctness nor in F-measure. According to our analysis, there is equivalence for correctness and for F-measure between individual and groupwise evaluation. For correctness even differences of up to 8% may safely be considered negligible, while for F-measure the threshold is 11%. Notice however that these thresholds are relatively large: it would mean, for example, that differences up to 11% false positives and false negatives are not meaningful; which, in some settings such as formal assessments of conformity, may leave too much ambiguity.

On the other hand, there is an effect for sensitivity (an increase of 8% for our sample). Correctness is affected by success criteria level (63% for A *vs* 45% for AA), sensitivity by method (57% *vs* 65%, for individual/groupwise evaluations) and by success criteria level (68% for A *vs* 48% for AA); F-measure varies accordingly, affected only by success criteria level (62% for A *vs* 50% for AA). These numbers tell us that level AA success criteria are more error prone and, because no interaction is detected, the evaluation method has the same effect for both levels (none for correctness, about 8% improvement for sensitivity).

The absence of effects for correctness and presence of an actual effect for sensitivity suggest that false negatives were reduced in groupwise evaluations, while the number of false positives leveled. This may indicate that the violations that were found individually – were these violations correct or not – were considered as-is also when reported groupwise if a second individual had initially reported an absence of violations, no matter if this absence was reported correctly. This interpretation is consistent with data shown by Table 9. Therefore, we can say that one of the strengths of groupwise evaluation is that it helps identifying those violations that were not found individually. However, there is no effect for correctness as while the number of true positives may increase a number of false positives are also introduced, neutralizing in this way any positive effect.

One implication of this phenomenon is that when carrying out groupwise evaluations, individuals should be explicitly encouraged to discuss how success criteria were rated, especially, if individuals agreed. One way for spurring the discussion would be by instructing them to elaborate and walkthrough on their rationale for their ratings. Instructions should also warn evaluators that more careful scrutiny should be given to possible false positives, even when members of the group agree on that outcome.

6.3. Reliability

Hypothesis **H3** states that evaluators working in groups compared to individuals achieve more reliable results in terms of significantly better max-agreement and any two agreements. Our data partially supports this – there is a significant difference on max-agreement (79 *vs* 89%, for IE and IEM respectively) but not for any-two agreement. Max-agreement depends on the success criteria levels, being higher for A (80% and 89%, for individual/groupwise evaluations) than for AA (77% *vs* 89%). When interpreting these results consider that the minimum possible value for max-agreement is 33%. Notice that complexity of success criteria has no effect on max-agreement.

For any-two agreement (the fraction of success criteria that any pair of evaluators/groups agreed upon) we see no difference due to the evaluation method (67% *vs* 72%); in fact they can be considered equivalent (up to a difference of 10%). No effect is due to success criteria level.

Thus, depending on how reliability is measured, a group effect can show up, with about a 10% increase when working in a group. Therefore, one should expect group assessments to be less variable when repeated over time than when the same kind of assessment are carried out individually.

Bear in mind, however, that our participants are quite similar in terms of background and experience; teaming people with more varied experience might lead to larger differences between groups.

6.4. Comparison with previous studies

In (Brajnik et al., 2012) effectiveness of novices and experts was compared when evaluating accessibility. That study is quite comparable to the current one as the demographic profile of participants and self-reported expertise is quite similar (see Table 12). Additionally, in this study we use a subset of the pages evaluated in that previous study.

Table 13 reproduces the values obtained by novices in (Brajnik et al., 2012) and allows us to compare the metrics between studies. When observing individual data (first two rows in Table 13), one of the conclusions is that the scores obtained for correctness, sensitivity and F-measure have been corroborated in this study. However, there is a sensible decrease when it comes to accuracy and an increase on max-agreement and any-two agreement. This divergence between studies may have been caused by the additional pages that were used in the previous experiment.

In (Brajnik et al., 2012) cumulative correctness was computed by adjoining the ratings produced by any pair of evaluators, for all possible pairs; similarly for sensitivity and F-measure. From Table 13 we can see that while correctness and F-measure of this "simulated" groupwise evaluations are similar to what we obtained this time, this is not true for sensitivity. In fact, when adjoining two evaluations performed totally independently, we obtained a sensitivity of 78%; on the other hand, letting evaluators interact reduces it to 65%. While there are no practical implications to be considered here (because sensitivity can only be assessed after-the-fact and only if one assumes that the correct ratings are already known), the results obtained from the current research suggest that letting people interact decreases sensitivity (by 13%). Thus, working in pairs improves sensitivity as opposed to working alone, but the improvement is smaller than expected.

6.5. Methodological considerations and future work

Regarding the validity of our findings, we argue that involving young students soon to become junior web developers with limited knowledge and skills in web accessibility, increases the ecological validity of this study, rather than undermining it (as usually happens when experiments involve university students). A survey run in Brazil on 2008 interviewed more than 600 web developers (Freire et al., 2008): they reported that formal training in web accessibility is limited, and lamented lack of experts. Similarly, a study with IBM developers (Trewin et al., 2010) reported that 60% of the respondents were accessibility novices or intermediate. We therefore argue that a large portion of people doing accessibility evaluations are not accessibility experts.

Notice also that although our sample is not particularly large, because it is homogeneous, because most of the measures have a relatively small range and because some of the effects were found to be statistically significantly different, the "wildcard" effect discussed in (Gray and Salzman, 1998, p. 210) does not hold, namely that "people that are significantly better or worse than average and whose performance in the conditions of the study do not reflect the [evaluation method] but reflect their wildcard status".

The following limits of our work should be considered when interpreting the data. First, we did not match people in groups so that they could build trust on each other before performing the assessment. It is likely that if members of the group knew each other before hand, the kind of intra-group agreement could be different. Secondly, our participants were very homogeneous in terms of knowledge and experience, reducing therefore the potential benefits of constituting a group. Varying these factors could lead to very different results and a stronger group effect. In particular, we argue that including in a group experienced people with different background (such as adding usability or user-experience experts to a team of accessibility specialists) could provide a strong benefit. Another way for varying the group members is to include people with disabilities as end-users.

Consider also that in the experiment we controlled only some of the relevant factors; future research avenues could aim at understanding:

- If the results we got hold also for groups larger than two.
- If the same kind of results could hold also for groups made of homogeneous but experienced accessibility evaluators.
- If members could focus on improving the previously written individual reports rather than, as we did,

Study	N	Age	Gender	Stimuli	Accessibility expertise	WCAG expertise
(Brajnik et al., 2012)	27	21–29, M=23, sd=1.9	4 female (15%)	Bloomberg (6), Facebook (7), IMDB (6), Sci.Am. (8)	Mdn=2, sd=0.72	Mdn=2, sd=0.71
this paper	20	21–33, M=22, sd=2.6	5 female (25%)	Bloomberg, IMDB, Sci.Am.,	Mdn=3, sd=0.52	Mdn=3, sd=0.9

Table 12. Comparison of participants demographics, expertise and web pages evaluated between previous and this study (N: number of participants).

Study	Acc.	C	S	F	A2A	MA
individually (Brajnik et al., 2012)	66	56	55	52	54	72
individually, this paper	59	56	56	53	67	79
2 users, cumulative (Brajnik et al., 2012)	–	54	78	around 60	–	–
groupwise, this paper	59	54	65	57	72	89

Table 13. Comparison of effectiveness scores (in percentage, Acc=accuracy, C=correctness, S=sensitivity, F=F-measure, A2A=any-2 agreement, MA=max-agreement) between the previous and this study.

asking group members to work on a single group report. Or, alternatively, if members could work directly on a group assessment, without carrying out a previous individual assessment.

Finally, research could be pursued aimed at finding out the reasons why certain success criteria were ranked high or low in terms of accuracy. All these issues point to possible interesting future research avenues that could pave the way to improve effectiveness and reliability of accessibility evaluation methods.

7. CONCLUSION

In this paper we presented the results of an experiment comparing performance of novice accessibility evaluators carrying out WCAG 2.0 conformance reviews to performance obtained when they work in teams of two.

Results indicate that accuracy of ratings is not significantly affected by group-work; similarly for correctness, F-measure and any-two agreement. Significant differences were found only for sensitivity (+8%) and max-agreement (+10%). Comparison of intra-group ratings shows that members of groups exhibited strong agreement among them, and with the group outcome. Therefore, overall, the conclusion is that constituting groups of two novice evaluators leads to a reduction of false negative rate (which goes as low as 38%) but no change in false positive rate, which remains high (about 45%). In addition, while working in a team of two improves the ability to catch all the true problems, the improvement is smaller than expected (by roughly 13%). Finally, when differences up to 8% in correctness and 11% in F-measure can be tolerated, the two accessibility evaluation methods are equivalent with respect to these performance indexes. Furthermore, groupwork is slightly more reliable (+10%).

The specific findings of this paper are the following:

- Groupwise evaluations produce the same overall accuracy although the accuracy of the worst and best success criteria tend to get closer to both ends.
- When carrying out groupwise evaluations, individuals should be explicitly encouraged to discuss how success criteria are rated no matter if individuals disagree or if they agree.
- Group assessments are less variable when repeated over time than when the same kind of assessment is carried out individually.
- Letting people interact reduces the false-negative rate.

From a practical viewpoint, this means that when reducing the false negative rate is a requirement, employing groups of two novice evaluators is more useful than asking a single one of them to perform the assessment. Therefore, especially in situations were developers are inexperienced in accessibility, working in teams could be beneficial. The benefits could increase if they are particularly focused on screening false positives. In general, we expect that teaming novices will lead to mutual learning and better understanding of the whole range of accessibility problems; however this might not significantly affect the accuracy of their assessments.

REFERENCES

S. Abou-Zahra. Web accessibility evaluation. In S. Harper and Y. Yesilada, editors, *Web Accessibility: A Foundation for Research*, Human-Computer Interaction Series, chapter 7, pages 79–106. Springer, London, first edition, Sept. 2008. ISBN 978-1-84800-049-0.

C. Bailey, Pearson E., and Gkatzidou V. Measuring and comparing the reliability of the structured walkthrough evaluation method with novices and experts. In *Web for All Conference*, Seoul, South Korea, April 2014. ACM, ACM Press.

G. Brajnik, Y. Yesilada, and S. Harper. The expertise effect on web accessibility evaluation methods. *Human-Computer Interaction*, 26(3):246–283, 2011. doi: 10.1080/07370024.2011.601670.

G. Brajnik, Y. Yesilada, and S. Harper. Is accessibility conformance an elusive property? a study of validity and reliability of WCAG 2.0. *ACM Trans. on Accessible Computing*, 2(4):8:1–8:28, March 2012. ISSN 1936-7228. doi: 10.1145/2141943.2141946. URL http://doi.acm.org/10.1145/2141943.2141946. doi: 10.1145/2141943.2141946.

DRC. The web: Access and inclusion for disabled people. Technical Report, Disability Rights Commission (DRC), UK, 2004.

L Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35(3):379–383, 2003. ISSN 0743-3808. doi: 10.3758/BF03195514. URL http://dx.doi.org/10.3758/BF03195514.

D. Fischer and T. Wyatt. The case for a WCAG-based evaluation scheme with a graded rating scale. In *RDWG Symposium on Website Accessibility Metrics*, W3C Research Notes. W3C, Dec. 2011 2012. URL http://www.w3.org/WAI/RD/2011/metrics/paper7/. Under publication.

A. Følstad. The effect of group discussions in usability inspection: A pilot study. In *Proc. of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, Nordichi '08, pages 467–470, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-704-9. doi: 10.1145/1463160.1463221. URL http://doi.acm.org/10.1145/1463160.1463221.

A. P. Freire, C. M. Russo, and R. P. M. Fortes. A survey on the accessibility awareness of people involved in web development projects in brazil. In *Proc. of the 2008 International Cross-disciplinary Conference on Web Accessibility (W4A)*, W4A '08, pages 87–96, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-153-8. doi: 10.1145/1368044.1368064. URL http://doi.acm.org/10.1145/1368044.1368064.

W.D. Gray and M.C. Salzman. Damaged merchandise: a review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, 13(3):203–261, 1998.

S.L. Henry and M. Grossnickle.
Just Ask: Integrating Accessibility Throughout Design.
Lulu.com, 2007.
ISBN 1430319526.

M. Hertzum and N.E. Jacobsen. The evaluator effect: a chilling fact about usability evaluation methods. *Int. Journal of Human-Computer Interaction*, 1(4):421–443, 2001.

M. Hertzum, N.Eb. Jacobsen, and R. Molich. Usability inspections by groups of specialists: Perceived agreement in spite of disparate observations. In *CHI 2002 Extended Abstracts on Human Factors in Computing Systems*, pages 662–663. ACM, ACM Press, 2002.

G.W. Hill. Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin*, 91(3):517–539, May 1982.

K. Hornbæk and E. Frøkjær. A study of the evaluator effect in usability testing. *Human-Computer Interaction*, 23(3):251–277, 2008.

W. Hwang and G. Salvendy. Number of people required for usability evaluation: the 10 ± 2 rule. *Communications of the ACM*, 53(5):130–133, May 2010. ISSN 0001-0782. doi: 10.1145/1735223.1735255. URL http://doi.acm.org/10.1145/1735223.1735255.

R. Jeffries, P.G. Polson, L. Razran, and M.E. Atwood. A process model for missionaries-cannibals and other river-crossing problems. *Cognitive Psychology*, 9(4):412 – 440, 1977. ISSN 0010-0285. doi: http://dx.doi.org/10.1016/0010-0285(77)90015-9. URL http://www.sciencedirect.com/science/article/pii/0010028577900159.

L.K. Michaelsen, W.E. Watson, and R.H. Black. A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology*, 74(5):834, 1989.

F.C. Miner. Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance*, 33(1):112 – 124, 1984. ISSN 0030-5073. doi: http://dx.doi.org/10.1016/0030-5073(84)90014-X. URL http://www.sciencedirect.com/science/article/pii/003050738490014X.

J. Nielsen. Finding usability problems through heuristic evaluation. In *Proc. of CHI 1992*, pages 373–380, Monterey, CA, USA, May 1992. ACM.

J. Nielsen. Heuristic evaluation. In J. Nielsen and R.L. Mack, editors, *Usability Inspection Methods.* John Wiley and Sons, 1994.

J. Nielsen. Heuristic evaluation. www.useit.com/papers/heuristic, 2002.

C. Power, A. Freire, H. Petrie, and D. Swallow. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proc. of the SIGCHI conference on human factors in computing systems, CHI 2012*, pages 433–442. ACM, 2012.

M. Schmettow. Sample size in usability studies. *Communications of the ACM*, 55(4):64–70, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133824. URL http://doi.acm.org/10.1145/2133806.2133824.

A. Sears. Heuristic walkthroughs: finding the problems without the noise. *Int. Journal of Human-Computer Interaction*, 9 (3):213–234, 1997.

J. Thatcher, M. Burks, C. Heilmann, S. Henry, A. Kirkpatrick, P. Lauke, B. Lawson, B. Regan, R. Rutter, M. Urban, and C. Waddell. *Web Accessibility: Web Standards and Regulatory Compliance*. FriendsofED, 2006.

S. Trewin, B. Cragun, C. Swart, J. Brezin, and J. Richards. Accessibility challenges and tool features: An IBM web developer perspective. In *Proc. of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, W4A '10, pages 32:1–32:10, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0045-2. doi: 10.1145/1805986.1806029. URL http://doi.acm.org/10.1145/1805986.1806029.

M. Vigo and G. Brajnik. Automatic web accessibility metrics: where we are and where we can go. *Interacting with Computers*, 23(2):137–155, March 2011. doi: doi:10.1016/j.intcom.2011.01.001.

W3C/WAI. Web content accessibility guidelines (wcag) 2.0. World Wide Web Consortium — Web Accessibility Initiative, www.w3.org/TR/WCAG20, December 2008.

W3C Web Accessibility Initiative WAI. Using combined expertise to evaluate web accessibility, 2002. URL http://www.w3.org/WAI/eval/reviewteams.html. Last checked: 15/01/2014.

W3C Web Accessibility Initiative WAI. Website accessibility conformance evaluation methodology (WCAG-EM) 1.0, 2014. URL http://www.w3.org/TR/WCAG-EM/.

W.E. Watson, L.K. Michaelsen, and W. Sharp. Member competence, group interaction, and group decision making: A longitudinal study. *Journal of Applied Psychology*, 76(6): 803, 1991.

Y. Yesilada, G. Brajnik, M. Vigo, and S. Harper. Exploring perceptions of web accessibility: a survey approach. *Behaviour and Information Technology*, 34(2):119–134, 2014. doi: 10.1080/0144929X.2013.8448238. URL http://dx.doi.org/10.1080/0144929X.2013.8448238.

A.1. DETAILED DATA

This appendix contains tables of data referred to in the body of the article.

SC	**fail**	**notapp**	**pass**	**Total**
2.4.2	0	0	20	20
2.3.1	0	7	13	20
1.4.2	0	26	0	26
2.2.1	1	19	0	20
1.2.2	1	27	0	28
3.2.1	2	0	26	28
2.2.2	3	16	1	20
1.2.3	3	25	0	28
2.1.1	5	0	15	20
2.4.3	5	0	15	20
2.1.2	6	0	14	20
1.4.1	6	1	13	20
1.3.3	6	2	11	19
1.3.2	6	2	12	20
3.3.1	8	12	0	20
2.4.4	9	0	11	20
3.2.2	9	1	10	20
1.2.1	9	31	0	40
3.1.1	13	4	3	20
2.4.1	14	3	3	20
1.3.1	15	2	11	28
4.1.1	16	0	4	20
4.1.2	17	1	2	20
3.3.2	18	5	6	29
1.1.1	20	0	0	20
Total	**192**	**184**	**190**	**566**

Table A1. List of level A success criteria and corresponding number of ratings that participants found.

SC	**fail**	**notapp**	**pass**	**Total**
3.2.3	0	15	5	20
2.4.5	1	4	15	20
1.2.4	1	19	0	20
3.3.4	1	19	0	20
3.2.4	3	10	7	20
1.2.5	3	25	0	28
1.4.4	9	0	27	36
1.4.5	9	1	20	30
2.4.6	10	0	18	28
3.3.3	10	10	0	20
1.4.3	16	0	1	17
2.4.7	21	0	19	40
3.1.2	21	13	6	40
Total	**105**	**116**	**118**	**339**

Table A2. List of level AA success criteria and corresponding number of ratings that participants found.

SC	fail	notapp	pass	Total
2.4.2	0	0	10	10
3.2.1	0	0	14	14
2.4.5	0	2	8	10
2.3.1	0	4	6	10
1.2.4	0	10	0	10
3.2.3	0	10	0	10
1.2.2	0	14	0	14
1.4.2	0	15	0	15
3.2.4	1	7	2	10
2.2.1	1	9	0	10
3.3.4	1	9	0	10
1.2.3	1	13	0	14
2.1.1	2	0	8	10
2.1.2	2	0	8	10
2.2.2	2	8	0	10
1.2.5	2	12	0	14
2.4.3	4	0	6	10
1.3.3	4	1	5	10
3.2.2	4	1	5	10
1.4.1	5	0	5	10
2.4.4	5	0	5	10
1.3.2	5	1	4	10
3.3.1	5	5	0	10
3.3.3	5	5	0	10
1.4.5	6	1	12	19
1.2.1	6	14	0	20
2.4.6	7	0	7	14
1.4.4	7	0	12	19
4.1.1	8	0	2	10
1.3.1	8	0	6	14
2.4.1	8	2	0	10
1.4.3	9	0	1	10
3.1.1	9	0	1	10
1.1.1	10	0	0	10
4.1.2	10	0	0	10
3.1.2	10	10	0	20
3.3.2	11	0	4	15
2.4.7	13	0	7	20
Total	**171**	**153**	**138**	**462**

Table A3. Success criteria (SC) and the number of ratings that the 10 groups produced.

SC	N	Tot	Accuracy
2.2.2	1	20	5.00
3.3.1	2	20	10.00
3.2.3	5	20	25.00
2.4.7	14	40	35.00
3.1.2	17	40	42.50
1.4.4	16	36	44.44
1.2.1	20	40	50.00
1.3.1	14	28	50.00
1.4.5	15	30	50.00
2.2.1	10	20	50.00
2.4.6	14	28	50.00
3.2.2	10	20	50.00
3.3.3	10	20	50.00
3.2.4	11	20	55.00
3.3.2	16	29	55.17
1.3.3	11	19	57.89
1.3.2	12	20	60.00
1.2.3	17	28	60.71
1.2.5	17	28	60.71
3.2.1	18	28	64.29
1.4.1	13	20	65.00
2.3.1	13	20	65.00
2.4.4	13	20	65.00
3.1.1	13	20	65.00
1.4.2	17	26	65.38
1.2.2	19	28	67.86
2.1.2	14	20	70.00
2.4.1	14	20	70.00
2.1.1	15	20	75.00
2.4.3	15	20	75.00
2.4.5	15	20	75.00
4.1.1	16	20	80.00
1.4.3	14	17	82.35
4.1.2	17	20	85.00
1.2.4	19	20	95.00
3.3.4	19	20	95.00
1.1.1	20	20	100.00
2.4.2	20	20	100.00

Table A4. Success criteria and corresponding accuracy values produced by individuals (Tot: number of ratings, N: number of correct ratings; Accuracy: accuracy in percentage).

SC	N	Tot	Accuracy
2.2.2	0	10	0.00
3.2.3	0	10	0.00
3.3.1	1	10	10.00
2.4.7	5	20	25.00
1.4.4	7	19	36.84
1.3.2	4	10	40.00
1.4.5	9	19	47.37
1.2.1	10	20	50.00
1.3.3	5	10	50.00
1.4.1	5	10	50.00
2.2.1	5	10	50.00
3.1.2	10	20	50.00
3.2.2	5	10	50.00
3.3.3	5	10	50.00
1.2.5	8	14	57.14
1.3.1	8	14	57.14
2.3.1	6	10	60.00
2.4.4	6	10	60.00
3.2.4	6	10	60.00
1.2.3	9	14	64.29
2.4.6	9	14	64.29
1.4.2	10	15	66.67
3.3.2	10	15	66.67
1.2.2	10	14	71.43
3.2.1	10	14	71.43
1.4.3	8	10	80.00
2.1.1	8	10	80.00
2.1.2	8	10	80.00
2.4.1	8	10	80.00
2.4.3	8	10	80.00
2.4.5	8	10	80.00
4.1.1	8	10	80.00
3.1.1	9	10	90.00
3.3.4	9	10	90.00
1.1.1	10	10	100.00
1.2.4	10	10	100.00
2.4.2	10	10	100.00
4.1.2	10	10	100.00

Table A5. Success criteria (SC) and corresponding accuracy values produced by groups (Tot: number of ratings, N: number of correct ratings; Accuracy: accuracy in percentage).