

CHAPTER FIVE

RESULTS, ANALYSIS AND DISCUSSION

5.1 INTRODUCTION

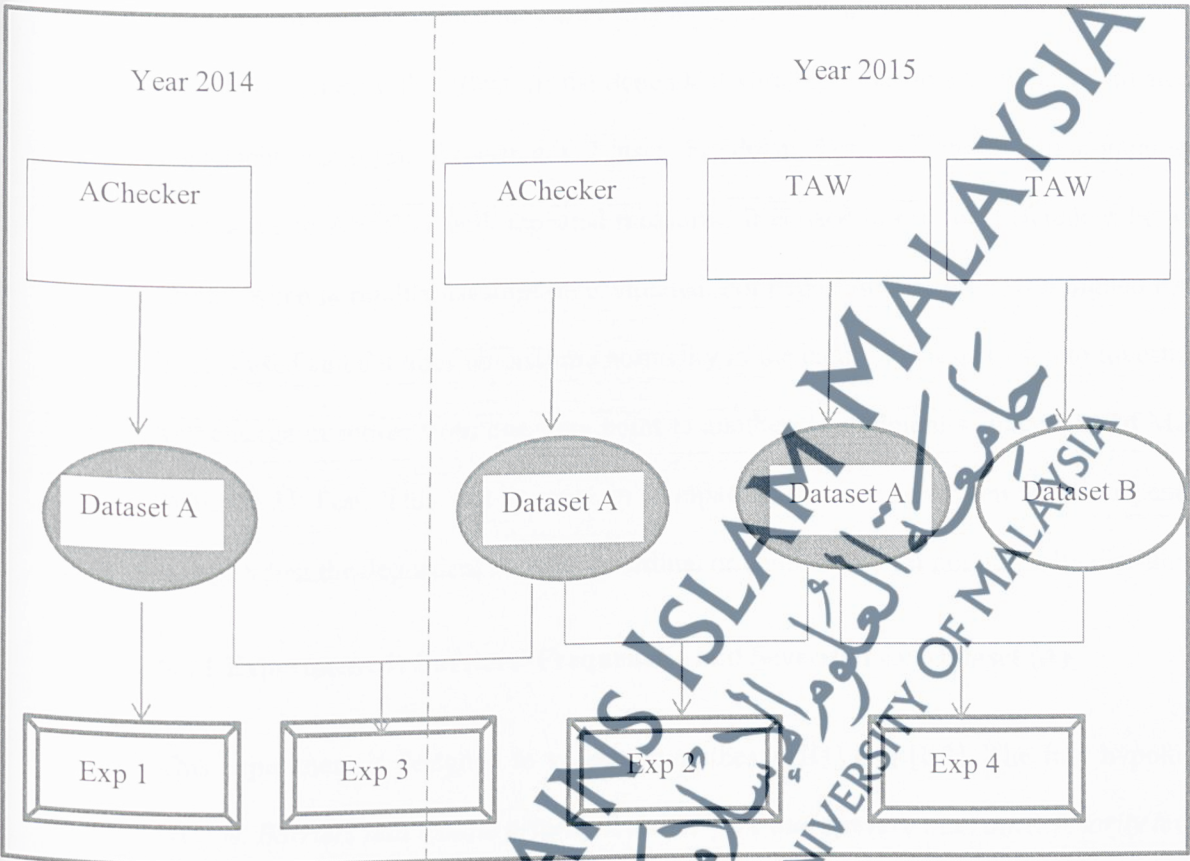
This chapter validates WABS metric quality in terms of validity, reliability and sensitivity. The results of quantitative analyses of the four experiments conducted to test the five hypotheses formulated in Chapter 3 are discussed. Moreover, the degree of accessibility level improvement achieved by the metric for a set of websites is demonstrated in section 5.5. Datasets (A) and (B) that were discussed in Chapter 3 were utilized to run the experiments using A-Checker and Taw automated tools. Dataset (A) contains high-traffic homepages of Malaysian websites, whereas dataset (B) contains top high-traffic global websites. Dataset (A) was used in the first three experiments and dataset (B) in the fourth experiment to analyze the effect of page sampling on barriers weight; when data are collected from different sample other than Malaysian websites. Known errors that do not need human intervention were collected from each homepage within these datasets. In other words, violations that can be checked automatically are considered, whereas generic problems and warnings that needed human review are ignored. In order to run the metric, it is needed to collect data about each error (barrier) frequency in each tested webpage using the aid of the automated tools. Unlike TAW tool, one shortcoming of A-Checker is the inability of the tool to list the frequency of each barrier. The generated report by the tool only shows an (x) notation next to the violated SC without mentioning its frequency. Therefore, a program was designed (as described in

section 3.4.1.2) to analyze the reports generated by A-Checker to count barriers' frequencies automatically which reduce the time needed for analyzing the results. Once the evaluation datasets and the automated tools were chosen, four experiments are designed to test the metric quality. The first experiment results are further explored to measure the degree of accessibility level improvement when the highly ranked severe barriers identified by WABS metric are removed. Descriptive and inferential statistics are adopted to analyze the experiments' results in order to validate the research hypotheses. Then, the quality of the metric is checked in terms of validity, sensitivity and reliability based on experiments findings. Table 5.1 summarizes the goal of each experiment, the year the experiment was conducted in, the automated tool was used, the dataset and the corresponding hypotheses. The experiments plan is illustrated in Figure 5.1

TABLE 5.1: Experiments’ Description

Experiment	Experiment 1	Experiment 2	Experiment 3	Experiment4
Goal	Measure the severity of each barrier	Check barriers' weight when evaluating by different tools	Check barriers' weight when evaluating over time by the same tool	Check barriers' weight when data are collected from different samples by the same tool
Year	2014	2015	2014/2015	2015
Automated Tools	A-Checker	A-Checker/ TAW	A-Checker	TAW
Dataset	“A”	“A”	“A”	“A” vs “B”
Hypothesis	H1: Barriers that violate priority level “A” are more severe than those that violate other priority levels. H2: Barriers that violate the same priority level have the same weight.	H3: Barriers' weight significantly changes when different tools are utilized	H4: Barriers' weight significantly changes over time	H5: Barriers' weight significantly changes when data are collected from different samples.

FIGURE 5.1: The Experiments Plan



5.2 RESULTS AND ANALYSIS FOR METRIC QUALITY

This section demonstrates barriers' weight generated by each experiment to show metric results under different contexts. The SPSS program version 22.0 (IBM Corp, 2013) is used to analyze quantitative data using descriptive and inferential statistics. Descriptive statistics are run on the first experiment to measure the central tendency of data using median. It is highly likely that the data are skewed, so using the median is the best choice since it is less affected by outliers and skewed data. Furthermore, inferential statistics are

run on the rest of experiments. Automated tools, time and samples are treated as independent variables since they are being manipulated and controlled within experiments in order to observe the effect on the dependent variables. Barriers weights are treated as dependent variables. Experiment 2 uses Friedman Test, which is a non-parametric alternative to ANOVA with repeated measures. It is used to test for differences between groups when normality assumption is violated. For experiment 3, Wilcoxon Signed Ranks Test is used since it does not assume normality in the data. This test is used to investigate any change in scores from one time point to another. Experiment 4 makes use of Mann-Whitney U Test. This test is used to compare differences between two independent groups when the dependent variable is ordinal or continuous, yet not normally distributed.

5.2.1 Experiment 1: Barriers' Frequencies and Severities for Dataset (A)

This experiment is designed to validate hypothesis [H1] and [H2]. The first hypothesis [H1] is: *Barriers that violate priority level "A" are more severe than other priority levels*, i.e. barriers that violate priority level A are more severe than barriers that violate priority levels AA and AAA. The second hypothesis [H2] is: *Barriers that violate the same priority level have the same weight*. To test the previous hypotheses, an experiment was conducted in October, 2014 to measure the severity of each barrier. A-Checker tool was adopted to work with dataset (A). Table 5.2 shows the total frequency of each barrier across the whole dataset in a descending order. For simplicity purposes, barrier id corresponds to SC id that has been violated. A-Checker is able to check 15 different types of barriers (barriers id). The most frequent barrier across the dataset (A) was scored by (1.1.1) with 4626 occurrences. In contrast, barrier (2.2.1) only appeared once.

TABLE 5.2: The Total Frequency of each barrier across Dataset (A)

Barrier ID	Total Frequency	Priority Level
1.1.1	4626(most frequent)	A
1.4.6	2327	AAA
2.1.1	710	A
1.4.4	592	AA
1.3.1	311	A
2.4.4	222	A
3.3.2	156	A
3.1.1	117	A
2.4.6	104	AA
1.4.1	76	A
4.1.1	52	A
2.2.2	35	A
2.4.2	16	A
2.4.1	15	A
2.2.1	1(least frequent)	A
Total	9360	

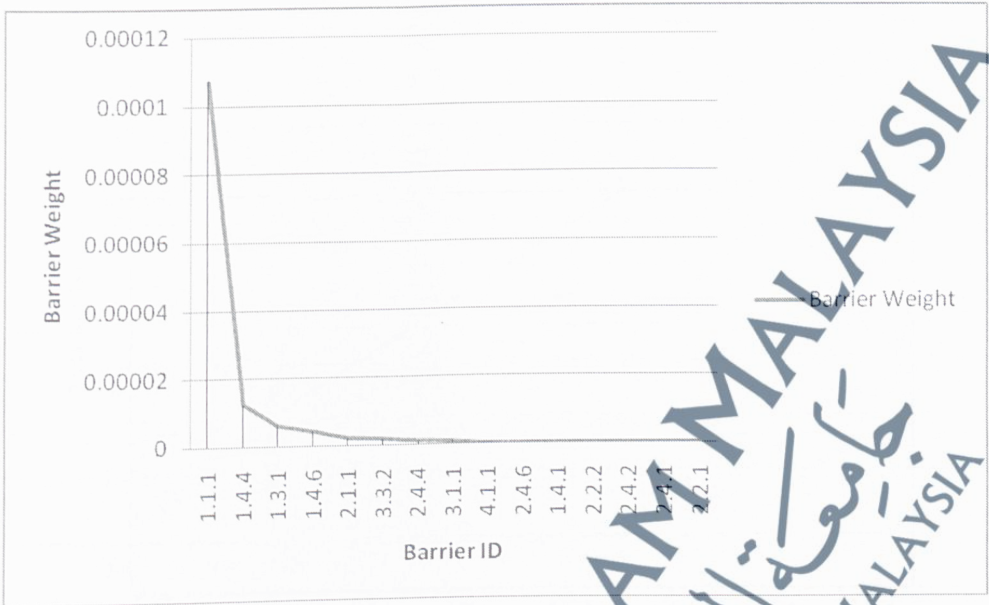
5.2.1.1 Experiment 1 results

After collecting the needed data, WABS metric was applied to calculate the weight of each barrier. Table 5.3 presents the weights generated by WABS for each barrier ordered from the most severe (highest weight) to the least severe barrier. The most severe barrier found in dataset “A” in 2014 was (1.1.1) while the least severe one was (2.2.1). Figure 5.2 depicts the weight of each barrier generated by WABS metric.

TABLE 5.3: The Weight for each Barrier in Dataset (A) using A-Checker

Barrier ID	Barrier Weight	Priority Level
1.1.1	0.000107(most severe)	A
1.4.4	1.22E-05	AA
1.3.1	6.14E-06	A
1.4.6	4.4E-06	AAA
2.1.1	2.04E-06	A
3.3.2	1.82E-06	A
2.4.4	1.05E-06	A
3.1.1	8.67E-07	A
4.1.1	1.35E-07	A
2.4.6	5.99E-08	AA
1.4.1	4.73E-08	A
2.2.2	1.73E-08	A
2.4.2	4.37E-09	A
2.4.1	3.81E-09	A
2.2.1	7.87E-11 (least severe)	A

FIGURE 5.2: Weights Generated by WABS Metric for each Barrier in Dataset (A)

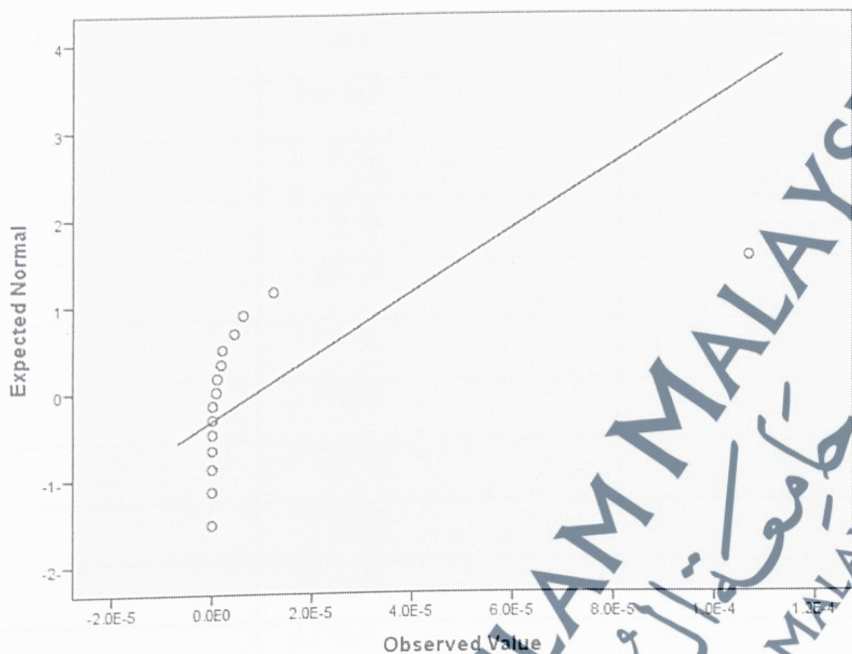


5.2.1.2 Analysis 1, testing data normality

The distributional shape of barriers weight in dataset (A) was examined through descriptive statistics to determine the extent to which the assumption of normality was met. Skewness (3.777, $SE= .580$), kurtosis (14.448, $SE= 1.121$), and the Shapiro-Wilk Test of normality ($S-W=.368$, $df= 15$, $p<.001$) suggests that normality is violated. Visually, Figure 5.3 displays departure from the diagonal line. Thus, the null hypothesis of normality is rejected and the weights appear to be positively skewed.

The average weights by barriers are not normally distributed. Hence, the Median and Quartiles provide more accurate measures of central tendency. The median weight per barrier by A-Checker ($N= 15$) for the year 2014 was $8.67E-7$ ($IQR= .0000044$). The maximum median weight was .0001070 scored by (1.1.1) and the minimum was $7.87E-11$ (2.2.1).

FIGURE 5.3: Normal Q-Q plot of Weights



5.2.1.3 Analysis 2, testing the relation between barrier frequency and its weight

From Tables 5.2 and 5.3, it can be noticed that barrier (1.1.1) was the most frequent and severe one at the same time, while (2.2.1) had the least frequency and severity in dataset (A). Consequently, further inferential analysis was carried out to investigate if there is a relation between barriers' frequency and severity. Table 5.4 shows each barrier along with its total frequency and weight.

A Spearman rank order correlation was conducted in order to determine if there were any relationships between barriers' frequency and weights. A two-tailed test of significance indicated that there was a significant positive relationship between frequencies and weights, $r_s(15) = .96$, $p < .001$. Therefore, the more frequent the barrier appears in the evaluation set, the highly likely it is to be more severe.

TABLE 5.4: Barriers with Corresponding Weight and Frequency

Barrier ID	Weight	Total Frequency
1.1.1	0.000107	4626
1.3.1	6.14E-06	311
1.4.1	4.73E-08	76
1.4.4	1.22E-05	592
1.4.6	4.40E-06	2327
2.1.1	2.04E-06	710
2.2.1	7.87E-11	1
2.2.2	1.73E-08	35
2.4.1	3.81E-09	15
2.4.2	4.37E-09	16
2.4.4	1.05E-06	222
2.4.6	5.99E-08	104
3.1.1	8.67E-07	117
3.3.2	1.82E-06	156
4.1.1	1.35E-07	52

5.2.2 Experiment 2: Comparing Barriers Ranking Against Tools

Experiment 2 tests the third hypothesis [H3] which is: *Barriers' weights significantly change when different tools are utilized*. The automated tool results may vary according to previous studies (Al-Khalifa et al., 2012; Abuaddous et al., 2013). Nonetheless, this does not necessarily mean the barrier ranking will be different since the metric takes into account different aspects (Anderson et al., 2000). In October, 2015, A-Checker and TAW

tools were utilized to evaluate dataset (A). The main goal of this experiment is to check if there is a significant difference between barriers' weights when different tools are applied to the same webpages and whether this difference affects the overall ranking of barriers. When A-Checker tool was run over dataset (A) in 2015, six websites could not be analyzed due to error type: "504 Gateway Time-out". TAW tool also faced some problems when analyzing three websites resulting in "Server error 500". Webpages that could not be evaluated were discarded.

5.2.2.1 Experiment 2 results

Table 5.5 summarizes the common barriers that could be detected by both A-Checker and TAW. None of the barriers that belong to priority "AA" and "AAA" was found to be in common. Table 5.6 shows barriers' ranking for both tools. The most severe barriers by A-Checker and TAW were (1.1.1) and (1.3.1), respectively. Barrier (2.2.1) was found to be the least severe barrier by both tools. Figure 5.4 illustrates A-Checker and TAW results.

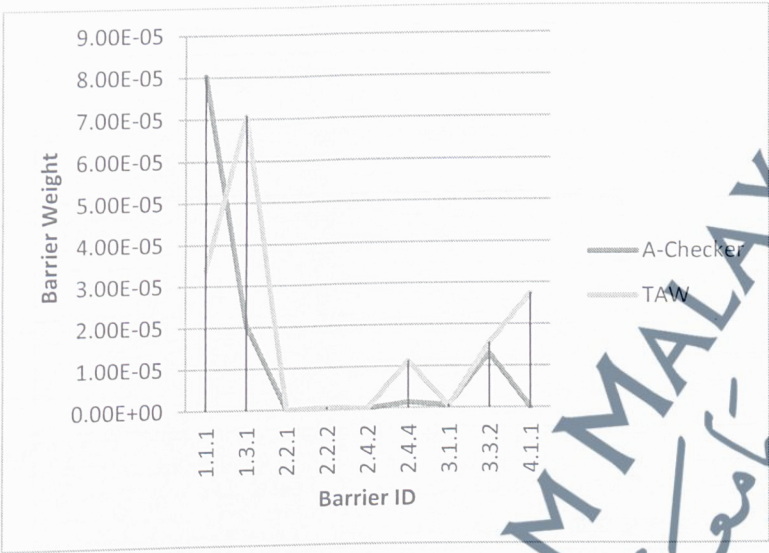
TABLE 5.5: Barriers' Weight Score when Utilizing A-Checker and TAW

Barrier ID	A-Checker	TAW	Priority Level
1.1.1	8.03E-05	3.4E-05	A
1.3.1	2.06E-05	7.08E-05	A
2.2.1	1.15E-10	2.95E-09	A
2.2.2	1.56E-08	4.88E-07	A
2.4.2	4.34E-09	1.85E-08	A
2.4.4	1.4E-06	1.12E-05	A
3.1.1	6.63E-07	4.81E-07	A
3.3.2	1.3E-05	1.55E-05	A
4.1.1	8.32E-08	2.74E-05	A

TABLE 5.6: Barriers' Ranking When Different Automated Tools are Applied to Dataset (A)

Barriers Ranking	A-Checker	TAW
1(most severe)	1.1.1	1.3.1
2	1.3.1	1.1.1
3	3.3.2	4.1.1
4	2.4.4	3.3.2
5	3.1.1	2.4.4
6	4.1.1	2.2.2
7	2.2.2	3.1.1
8	2.4.2	2.4.2
9(least severe)	2.2.1	2.2.1

FIGURE 5.4: Barriers' Weight Differences between Automated Tools

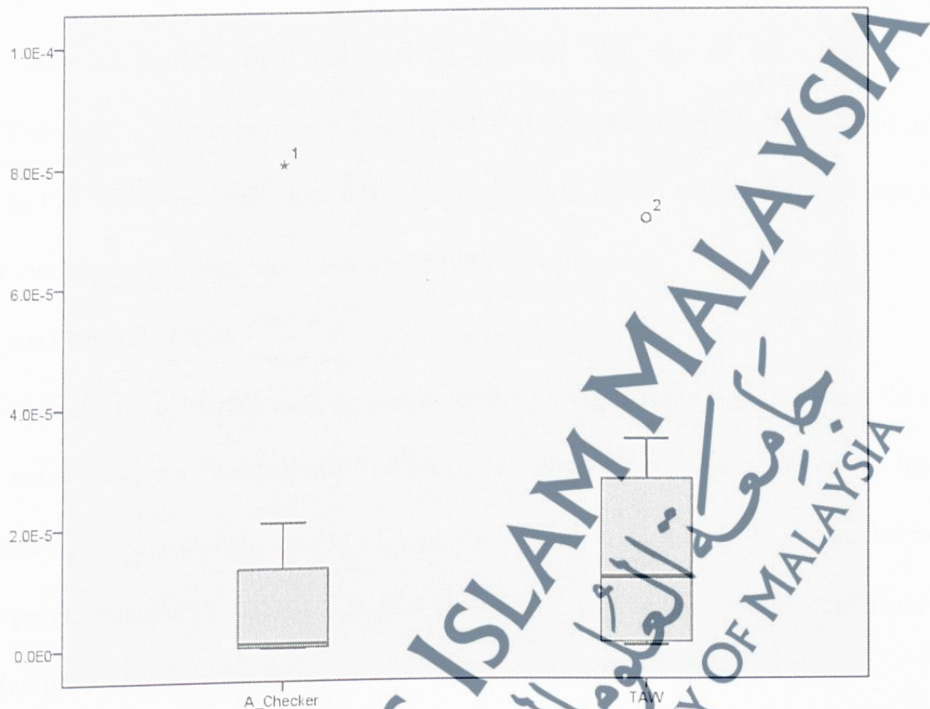


5.2.2.2 Experiment 2 analysis

A non-parametric Friedman test of differences among repeated measures was conducted to see if there was any significant change between barriers' weight when generated by A-Checker and TAW automated tools. At the (p-value =0.96) level of significance, there exists enough evidence to conclude that there was no statistically significant difference in barriers' weights when different tools were utilized. $\chi^2 (df=1, N=9) = 2.778$.

Box plot Figure 5.5 provides a summary of Tools' weight results. Since half the scores in a distribution are between the hinges, it is noted that half A-checker weights are between 1E-08 and 1.68E-05, whereas half TAW weights are between 2.5E-07 and 3.07E-05.

FIGURE 5.5: Box and whisker Plots for Weights Generated when using different Automated Tools



5.2.3 Experiment 3: Comparison of Barriers' Ranking over Time

The hypothesis that this experiment aims to check is the fourth hypothesis [H4] which is: *Barriers' weight significantly changes over time*. The Web nature is dynamic and so is the content. When conducting several evaluations for a specific dataset over a different time span, the results can change and this will affect the barriers' weight. This does not inevitably mean that the overall barriers' severity ranking will change (Anderson et al., 2000). This experiment aims to observe if there is significant change in barriers' weight and ranking when holding experiments in different intervals of time using the same automated tool and same dataset. Experiments 1 and 2 have tested dataset (A) in 2014 and 2015 by A-Checker tool. Thus, this experiment utilizes A-Checker results from

experiments 1 and 2 and compares them to observe the same dataset over time. Webpages that are not active anymore are excluded from this experiment. Six webpages were discarded since A-Checker could not analyze them in 2015 due to error type: “504 Gateway Time-out”. This experiment could give a slight indication of the accessibility condition in the Malaysian websites. Also, the evaluation would reveal if the designers' awareness increases over time when designing their websites.

5.2.3.1 Experiment 3 results

Table 5.7 displays each barrier with its corresponding weight scored in 2014 and 2015. Table 5.8 shows barriers' ranking in 2014 and 2015 ordered from most severe to least severe based on their weights. Barrier (1.1.1) was found to be the most severe barrier across the years, whereas (2.2.2), (2.4.2), (2.4.1) and (2.2.1) scored the least weights over the indicated time.

TABLE 5.7: Barriers' Weight Scores in 2014 and 2015

Barrier Weight	2014	2015	Priority Level
1.1.1	0.000107	8.03E-05	A
1.3.1	6.14E-06	2.06E-05	A
1.4.1	4.73E-08	2.38E-06	A
1.4.4	1.22E-05	5.19E-06	AA
1.4.6	4.40E-06	4.19E-06	AAA
2.1.1	2.04E-06	1.79E-06	A
2.2.1	7.87E-11	1.15E-10	A
2.2.2	1.73E-08	1.56E-08	A
2.4.1	3.81E-09	2.34E-09	A
2.4.2	4.37E-09	4.34E-09	A
2.4.4	1.05E-06	1.40E-06	A
2.4.6	5.99E-08	1.92E-07	AA
3.1.1	8.67E-07	6.63E-07	A
3.3.2	1.82E-06	1.30E-05	A
4.1.1	1.35E-07	8.32E-08	A

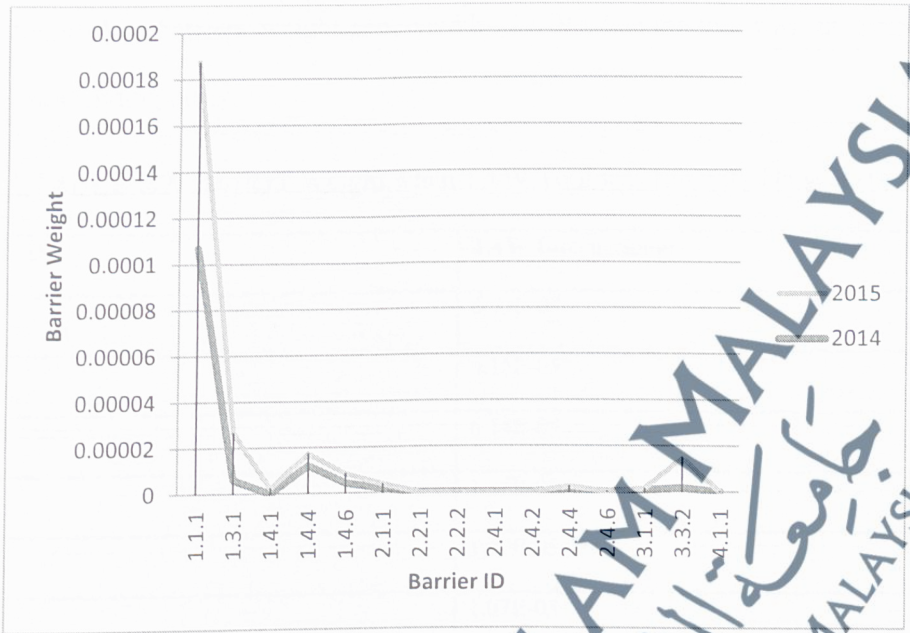
TABLE 5.8: Barriers' Ranking for Dataset (A) in 2014 and 2015

Barrier Ranking	2014	2015
1(most severe)	1.1.1	1.1.1
2	1.4.4	1.3.1
3	1.3.1	3.3.2
4	1.4.6	1.4.4
5	2.1.1	1.4.6
6	3.3.2	1.4.1
7	2.4.4	2.1.1
8	3.1.1	2.4.4
9	4.1.1	3.1.1
10	2.4.6	2.4.6
11	1.4.1	4.1.1
12	2.2.2	2.2.2
13	2.4.2	2.4.2
14	2.4.1	2.4.1
15(least severe)	2.2.1	2.2.1

5.2.3.2 Experiment 3 analysis

In order to test whether barriers weight changed significantly across the years, a Wilcoxon signed-rank test was run. The medians for 2014 and 2015 were 8.67E-07 and 1.4E-06, respectively. The Wilcoxon signed-rank test showed that barriers weight did not elicit a statistically significant change between 2014 and 2015 ($Z = -.227, p = .820$). Consider Figure 5.6, which plots barriers weight over time.

FIGURE 5.6: Barriers' Weight Differences over the Years



5.2.4 Experiment 4: Comparison of Barriers' Ranking with Different Samples

This experiment validated hypothesis five [H5]: *Barriers weight significantly changes when data are collected from different samples*. Conducting experiments with webpages collected from different samples could affect barriers' weight. Notwithstanding, this does not mean that the overall ranking of barriers will change (Anderson et al., 2000). Experiment 4 is intended to investigate the effect of collecting data from different samples. In October, 2015 an experiment using TAW tool against dataset (B) was carried out and then compared to TAW results for dataset (A) from experiment 2. TAW can check 16 different types of barriers. The experiments would discover the common programming practices of Malaysian Web developers concerning web accessibility. This experiment can show if the metric results could be generalized. Subsequently, WABS metric was applied to the results to calculate each barrier's weight.

5.2.4.1 Experiment 4 results

Table 5.9 displays barriers' weight generated by TAW when the tool was run over dataset (B) in a descending order.

TABLE 5.9: Barriers' Weight when TAW Tool was run with Dataset (B)

Barrier ID	TAW International
1.1.1	6.73E-05
1.3.1	9.15E-05
4.1.1	6.14E-05
4.1.2	1.35E-05
3.2.2	1.07E-05
3.3.2	1.07E-05
2.4.9	7.47E-06
3.1.1	1.17E-06
2.4.10	1.02E-06
2.1.3	6.44E-07
2.4.2	2.17E-08
2.2.1	8.64E-09
2.4.4	1.09E-09
3.2.5	9.65E-10
2.2.2	1.84E-10
2.2.4	1.84E-10

TAW results from experiment 2 with dataset (A) were compared with those from dataset (B). Table 5.10 shows each barrier with corresponding weight generated from datasets (A) and (B). Barriers ranking can be found in Table 5.11 ordered from most severe to

least severe barrier. Barriers (1.3.1), (1.1.1) and (4.1.1) scored the highest weights in both groups.

TABLE 5.10: Barriers with their Corresponding Weights in Datasets (A) and (B)

Barrier ID	TAW Malaysia (A)	TAW Global (B)	Priority Level
1.1.1	3.40E-05	6.73E-05	A
1.3.1	7.08E-05	9.15E-05	A
2.1.3	7.01E-07	6.44E-07	AAA
2.2.1	2.95E-09	8.64E-09	A
2.2.2	4.88E-07	1.84E-10	A
2.2.4	8.26E-10	1.84E-10	AAA
2.4.2	1.85E-08	2.17E-08	A
2.4.4	1.12E-05	1.09E-09	A
2.4.9	4.28E-06	7.47E-06	AAA
2.4.10	1.28E-06	1.02E-06	AAA
3.1.1	4.81E-07	1.17E-06	A
3.2.2	1.69E-07	1.07E-05	A
3.2.5	1.57E-08	9.65E-10	AAA
3.3.2	1.55E-05	1.07E-05	A
4.1.1	2.74E-05	6.14E-05	A
4.1.2	2.08E-05	1.35E-05	A

TABLE 5.11: Barriers' Ranking when Different Datasets are Used

Barrier Ranking	TAW Malaysia (A)	TAW Global (B)
1(most severe)	1.3.1	1.3.1
2	1.1.1	1.1.1
3	4.1.1	4.1.1
4	4.1.2	2.4.4
5	3.3.2	4.1.2
6	2.4.4	3.3.2
7	2.4.9	2.4.9
8	2.4.10	3.1.1
9	2.1.3	2.4.10
10	2.2.2	2.1.3
11	3.1.1	3.2.2
12	3.2.2	2.4.2
13	2.4.2	2.2.2
14	3.2.5	2.2.4
15	2.2.1	3.2.5
16(least severe)	2.2.4	2.2.2

5.2.4.2 Experiment 4 analysis

A Mann-Whitney U test was conducted to check whether there was significant change when data are collected from different samples. The Mann-Whitney U analysis sorted the data by order and assigned a rank to each number. The ranks were compared between the two groups (Datasets A and B) and summed up to determine if any difference exists. The results of the test were insignificant $U=126$, ($Z = -.075$, $p=.940$) and the difference

between groups was small ($r = -.01$). Dataset A (Malaysia) had an average mean rank of 16.38, while dataset B (Global) had an average mean rank of 16.63. For visual representation of the differences between barriers' weights when different dataset are used, see Figure 5.7.

FIGURE 5.7: Barriers' Weight Differences between Datasets



5.3 DISCUSSION

The purpose of this section is to summarize and comment on the results of the four experiments. The hypotheses formulated previously to answer the third research question are also discussed. Additionally, the experiments' analyses reported in this chapter are specifically discussed in this section in order to examine WABS metric quality.

5.3.1 Discussion on Experiment 1

The research sub question addressed by this experiment is the following: To what extent does each barrier contribute to accessibility problems in a given evaluation set? What are

the severity and the weight of each barrier when evaluating a specific dataset? Do the barriers which violate higher priority, score higher weight than the others (more severe)? Does barrier ranking depend only on the priority level?

Two hypotheses were formulated to answer this research question. The first hypothesis [H1] was: Barriers that violate priority level “A” are more severe than the others. The second hypothesis [H2] was: Barriers that violate the same priority level have the same weight. Before analyzing the results of the first experiment, normality was tested for dataset (A) to choose appropriate tests for each experiment. The number of tested barriers was slightly small, thus the data was skewed and violated normality assumption by Shapiro-Wilk Test.

Table 5.3 shows the unique weight scored by each accessibility barrier. It can be noticed that barriers that belong to higher priority level are not necessarily more severe than other types of barriers. For example, barriers (1.4.6) and (1.4.4) which belong to priority levels “AAA” and “AA” respectively were more severe than other barriers that belong to level “A”. Accordingly, [H1] is rejected.

Table 5.3 also shows how barriers weights that belong to the same priority level are different. For example, barriers (1.1.1), (3.3.2) and (2.2.1) scored different weights even though they belong to priority “A”. Thus, [H2] is rejected.

This experiment confirms that barriers weight does not merely depend on priority level. The total frequency of a barrier in the dataset has a large effect on its weight. Spearman

rank correlation coefficient showed evidence of a positive relationship between barrier frequency and weight with $r_s(15) = .96$ and $p < .001$.

5.3.2 Discussion on Experiment 2

The research sub question addressed by this experiment was the following: Do barriers' weights significantly change when different automated tools are utilized? Are the metric's results, tool independent? If there is a significant change, to what extent does it affect the overall ranking? The hypothesis stated a priori to answer this question was [H3]: Barriers weight significantly changes when different tools are utilized.

Since the data are not normally distributed, a Friedman test was run to measure if there is a significant change in barriers' weight when different tools are utilized. The results showed insignificant change at the ($p\text{-value} = .096$) level of significance. Hence, [H3] hypothesis is rejected. Nonetheless, barriers' ranking, slightly changed when the experiment was run with different tools. When A-Checker was used, the most severe barrier was (1.1.1), followed by (1.3.1). However, when TAW was utilized, (1.3.1) had the highest weight, followed by (1.1.1), as shown in Table 5.6. The slight differences can be due to different results produced by different tools as indicated by previous studies (Al-Khalifa et al., 2012; Abuaddous et al., 2013).

5.3.3 Discussion on Experiment 3

The research sub question addressed by this experiment is the following: Is there a significant change in barrier weight when evaluating the same dataset over time? If yes, does this change affect the overall ranking? The hypothesis based on this question was [H4]: Barriers' weight significantly changes over time

Table 5.8 demonstrates barriers ranking over the years. It can be observed that in 2014, barrier (1.1.1) was the most severe, followed by (1.4.4). In 2015, (1.1.1) remained in the first place, followed by (1.3.1) in the second place. The four last barriers stayed the same for both cases. One possible explanation of the small changes in barriers' ranking is the dynamic nature of the Web, as webpages contents change over time resulting in new barriers and different frequencies. However, Wilcoxon signed-rank test did not indicate a statistically significant difference between barriers' weight over time ($Z = -.227, p = .820$). Thus, [H4] is rejected.

5.3.4 Discussion on Experiment 4

The research question addressed by this experiment is the following: Does barriers' weight significantly change when data are collected from different samples? Can the results of the metric be generalized or are they bound to each experiment? The hypothesis formulated to answer this research question was [H5]: Barriers' weight significantly changes when data are collected from different samples.

A Mann-Whitney U test was conducted to test if there was a significant change between barriers' weight when data are collected from different samples. The results of the test were insignificant $U=126, (Z = -.075, p = .940)$ and the difference between groups was small ($r = -.01$). As shown in Table 5.11 above, the first three barriers had the same ranks in both groups, while the ranking for the rest of the barriers did not significantly change, so [H5] is rejected. The small changes in barriers ranking can be due to different programming practices followed by Web developers when coding their webpages. To put

it another way, some companies compel Web developers to adhere to their accessibility policy while others do not have this kind of regulations.

5.4 DISCUSSION ON METRIC QUALITY

This work focuses on the effectiveness of the metric which refined into validity, sensitivity and reliability.

5.4.1 Validity

The theoretical validation confirms that WABS metric was not built arbitrarily since it adheres to the requirements and assumptions made in Chapter 4 and follows proved principles found in Information Retrieval and Mathematics. The requirements to build a good metric suggested by the eminent researchers (Zeng, 2004; Vigo et al., 2007) were carefully followed when formulating variables, constants and final metric. Furthermore, Web accessibility metrics proposed by different researchers (Sullivan & Matson, 2000; Zeng, 2004; Arrue et al., 2005, Velleman et al., 2006) have been extensively studied from the viewpoint of formulations and goals. WABS metric utilizes some parameters used by metric proposed by (Zeng, 2004). The metric formulation in terms of barriers weighting and ranking is inspired by the vector space model. Thus, the metric measurements cover:

- (1) The importance of the barrier to the other barriers that violate the same priority level;
- (2) The importance of the barrier to the webpage; and (3) The importance of the barrier to all the other barriers in the whole dataset (all websites). Two principals were used from vector space model, namely the extension of Pythagorean Theorem (Weisstein, 2014), which is used to measure barrier length in different context (this principle covers the first

and third measurements). The second principle is TF-IDF (Sparck Jones, 1972) which was adjusted to meet the second measurement.

For empirical validation, the first experiment was conducted and analyzed to check if the metric values have discriminating power and if the barrier frequency has a positive correlation with its weight. Table 5.3 shows that the metric results were positive, finite and normalized with continuous range of values fell within the range (0-1) as intended by the specification. It is obvious that metric values have large discriminating power which supports Sheppard & Darrel (1993) contention. Table 5.3 also demonstrates how barriers weights that belong to same priority level are different. For example, barriers (1.1.1), (4.1.1) and (2.4.1) scored different weights even they belong to priority "A". Each barrier has a unique value even for barriers belong to the same priority level. This allows observing the influence of each barrier that belongs to the same priority level.

Moreover, the empirical validity of the metric highlighted in this chapter was tested by the correlation analysis between barrier frequency and barrier weight. Spearman rank correlation in subsection (5.2.1.3) showed that there is a positive correlation between the total number of accessibility barriers and the weight obtained for each barrier with $r_s(15) = .96$ and $p < .001$. This test concludes that barriers' weight does not only depend on priority level. The total frequency of a barrier in the dataset has a large effect on its weight. This confirms the logical assumption that the more the barrier appears in a dataset, the highly likely the barrier will be more severe. Therefore, when WABS metric was formulated it took into account the barrier frequency measure.

5.4.2 Sensitivity

The accuracy of metric measurements was designed and greatly enhanced in order to be able to discriminate between accessibility barriers. Experiments' results have indicated the ability of the metric to discriminate between accessibility barriers' severity by giving a positive numerical score based on the following measurements: (1) The importance of the barrier to other barriers that violate the same priority level; (2) The importance of the barrier to the webpage; and (3) The importance of the barrier to the whole dataset. The conducted experiments have demonstrated the sensitivity of the proposed metric. However, the metric was not too sensitive to small changes in websites contents when the experiments were conducted in different contexts (time span, tools and samples) since extremely sensitive metrics can over-react to small changes in web contents. Of greater concern, is the fact that highly dynamic websites can result in new contents which could affect metric results. The previous experiments showed a change, albeit insignificant, in accessibility barriers weight.

5.4.3 Reliability

Three experiments were conducted to examine metric behavior under different contexts:

- 1) One to study the differences in the metric scores when metric is fed with data produced by different automated tools.
- 2) A second to see how reliability changes when metric is carried out at different times.
- 3) A third to analyze the effects of page sampling when webpages are collected from different samples.

The proposed metric shows reliability, because all the measurements are reproducible and consistent when calculated by different evaluators under similar circumstances (same tool, sample and time interval). Moreover, the differences between metric scores were not significant when the metric was applied in different contexts. Traditional reliability measurements such as inter-rater and internal consistency are not applicable to WABS metric. However, small changes of metric scores were observed under different conditions. This can be due to the dynamic nature of websites which could result into different contents with different number and type of accessibility errors. Furthermore, WABS metric utilizes error reports generated by automated tools; it does not involve any subjective judgment. The reliability of results could rely on the automated tool being used.

5.5 RESULTS AND ANALYSIS FOR MEASURING THE IMPROVEMENT OF ACCESSIBILITY LEVEL

The first experiment of this work was conducted in October 2014 using A-Checker tool and dataset (A). Zeng's modified metric (formula 3.3) was applied to A-Checker's reports for each website in dataset (A) to calculate the accessibility score for each website. Table 5.12 shows the accessibility score for a twenty websites of dataset (A) using formula (3.3) in a descending order. These websites represents 20 Public Higher Educational Institutions in Malaysia. Higher score means low accessibility level and zero score means the website has not violated any accessibility success criteria and has no barriers. The full list of results can be found in Appendix A.

TABLE 5.12: Web Accessibility level for 20 Public Universities in Malaysia before Fixing Top Severe Barriers

Rank	Website Name	University Website	Web Accessibility score before fixing top severe barriers
1	upnm	http://www.upnm.edu.my/en/	77.28 (lowest Accessibility)
2	unisza	http://www.unisza.edu.my/index.php?lang=en	32.8
3	umk	http://www.umk.edu.my/rasmi/	30.64
4	uum	http://www.uum.edu.my/index.php/en/	24.16
5	ukm	http://www.ukm.my/v6/	22.48
6	upsi	http://www.upsi.edu.my/	22.2
7	um	http://www.um.edu.my/	19.64
8	upm	http://www.upm.edu.my/?LANG=en	13.44
9	umt	http://www.umd.edu.my/	11.84
10	utm	http://www.utm.my/	10.6
11	usm	http://www.usm.my/index.php/en/	10.32
12	iium	http://www.iium.edu.my/	5.72
13	ump	http://www.ump.edu.my/en	4.88
14	unimas	http://www.unimas.my/index.php/en/	2.24
15	uthm	http://www.uthm.edu.my/v2/	1.04
16	utem	http://www.utem.edu.my/web2012/	0.8
17	usim	http://www.usim.edu.my/en/index.php	0.44
18	uitm	http://www.uitm.edu.my/index.php/en/	0
19	ums	http://www.ums.edu.my/v5/	0
20	unimap	http://www.unimap.edu.my/	0

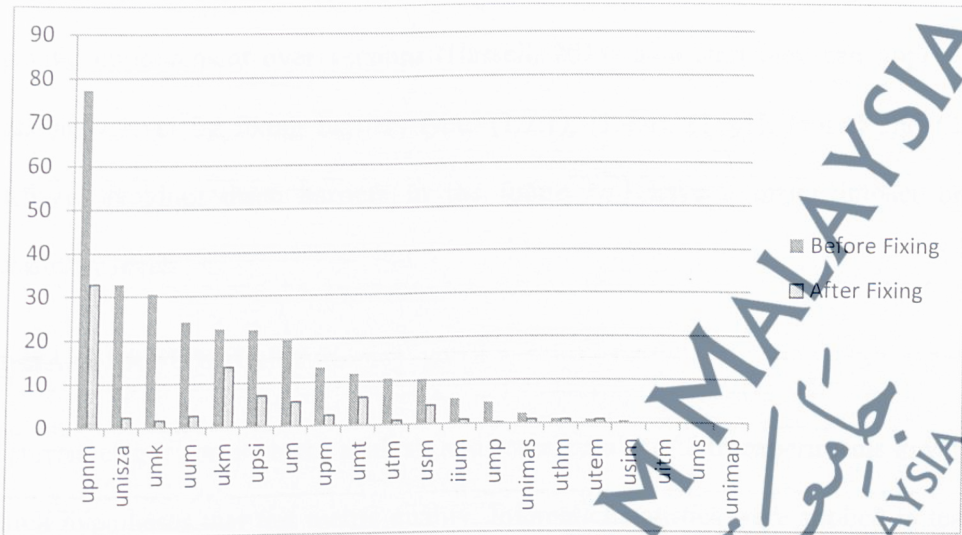
As illustrated earlier in Table 5.3, 15 barriers id (barriers type) can be identified using A-Checker. WABS metric assigned a weight for each of the 15 barrier types based on the severity and they were ranked accordingly. Some websites of the dataset found to contain all the types of barriers, while others have some, except for websites that were barrier free. The researcher chose to remove top five severe types of barriers which were identified by WABS. The researcher removed barriers to simulate fixing them since she is not authorized to fix them. Based on Table 5.3, the top five highly ranked barriers are (1.1.1, 1.4.4, 1.3.1, 1.4.6, 2.1.1), respectively. Once these barriers were removed from each website in dataset (A), the accessibility score for each website was measured again using formula (3.3). Table 5.13 and Figure 5.8 show the accessibility score of websites before and after removing highly ranked barriers. As can be noticed from Table 5.13

some websites scored zero before fixing severe barriers. Thus, these websites will not be useful for the comparison since they have no barriers to be fixed. Besides, some websites did not exhibit any change in accessibility score after removing the top five severe barriers like website 16 (Utem). This means that these websites did not contain any of the five severe barriers before and thus once these barriers were removed from the dataset they did not get affected and the score remained the same.

TABLE 5.13: The Difference between Web Accessibility Score for 20 Public Universities in Malaysia before and after Fixing Top Severe Barriers

Rank	University Name	Web Accessibility score before fixing top severe barriers	Web Accessibility score after fixing top severe barriers
1	upnm	77.28	32.8
2	unisza	32.8	2.4
3	umk	30.64	1.6
4	uum	24.16	2.56
5	ukm	22.48	13.76
6	upsi	22.2	7.2
7	um	19.64	5.6
8	upm	13.44	2.4
9	umt	11.84	6.4
10	utm	10.6	0.96
11	usm	10.32	4.32
12	iium	5.72	0.8
13	ump	4.88	0.16
14	unimas	2.24	0.8
15	uthm	1.04	0.8
16	utem	0.8	0.8
17	usim	0.44	0
18	uitm	0	0
19	ums	0	0
20	unimap	0	0

FIGURE 5.8: Web Accessibility Score Differences for 20 Public Universities in Malaysia before and after Fixing Top Severe Barriers



Among 500 websites of dataset (A), 28 websites scored zero and were barrier free. So, they were excluded from the comparison and 472 websites were compared.

A paired t-test was run on a sample of 472 websites to determine whether there was a statistically significant (Mean) difference between the accessibility level before and after fixing severe barriers. Since the higher the score obtained by formula (3.3) the lower the accessibility level, a higher Mean (M) indicates lower accessibility level. The difference in average number of accessibility level score between the two groups was statistically significant ($t = 8.467, p \approx 0.001 < 0.05 = \alpha$). After removing the top high five barriers a better accessibility level was obtained ($M=4.18, SE=5.36$) as opposed to before removing barriers ($M=56.25, SE=135$); a statistically significant accessibility level increase of 52.0644 (95% CI, 64.1475 to 39.9813) score. High score means lower accessibility and drop of score from 64.1475 to 39.9813 means the accessibility level has significantly improved.

This comparison shows how fixing top severe barriers identified by WABS metric has a big influence on accessibility level. Since Web developers aim for a continuous, pragmatic improvement over versions (Hassell, 2013) as a start they can improve the accessibility level by fixing barriers type: (1.1.1), (1.4.4), (1.3.1), (1.4.6) and (2.1.1). Moreover, avoiding these barriers in the future will have a major impact on the accessibility level.

5.6 CHAPTER SUMMARY

The current chapter presents the analysis and actual results of four experiments to validate research hypotheses that test metric quality. Inferential statistics were applied to test the experiments' hypotheses. The first experiment aims to test two hypotheses. The first hypothesis is [H1]: Barriers that violate priority level "A" are more severe than those that violate other priority levels. Meanwhile, the second hypothesis is [H2]: Barriers that violate the same priority level have the same weight. The results of the first experiment confirm that barriers weight does not merely depend on priority level and the total frequency of a barrier in the dataset has a large effect on its weight. Spearman rank correlation coefficient showed evidence of a positive relationship between barrier frequency and weight with $r_s(15) = .96$ and $p < .001$. Thus, hypothesis [H1] and [H2] are rejected. The second experiment compares barriers ranking against tools to validate the third hypothesis which is [H3]: Barriers' weight significantly changes when different tools are utilized. The results of this experiment shows that at the (p -value = .096) level of significance, there exists enough evidence to conclude that there was no statistically significant difference in barriers' weights when different tools were utilized χ^2 ($df=1$,

$N=9) = 2.778$, accordingly [H3] is rejected. The third experiment compares barriers ranking over time in order to test the fifth hypothesis which is [H4]: Barriers' weight significantly changes over time. The Wilcoxon signed-rank test showed that barriers weight did not elicit a statistically significant change between 2014 and 2015 ($Z = -.227$, $p = .820$), thus [H4] is rejected. The fourth experiment compares barriers ranking when samples are collected from Malaysian websites and from global websites. This experiment aims to check the final hypothesis which is [H5]: Barriers' weight significantly changes when data are collected from different samples. Mann-Whitney U analysis results were insignificant $U=126$, ($Z = -.075$, $p=.940$) and the difference between groups was small ($r = -.01$), so [H5] is rejected. Since all the hypotheses are rejected, WABS metric proved to meet the requirements which a valid and reliable metric should have. The results of first experiment were further explored and analyzed by paired samples t-test to compare the degree of accessibility level improvement for dataset (A) before and after removing the highest five severe barriers identified by WABS metric. At the ($t = 8.467$, $p \approx 0.001 < 0.05 = \alpha$) level of significance there exists enough evidence to conclude that there was a statistically significant improvement of accessibility level after removing the highest five severe barriers.