ENTER2006 - 13th International Conference on Information Technology and Travel & Tourism
January 18th – 20th, 2006 - Lausanne, Switzerland

In Hitz, M., Sigala, M., Murphy, J. (Eds.), Information and Communication Technologies in Tourism 2006 - Proceedings of the International Conference in Lausanne - Switzerland, Springer, Wien, 2006, pp. 320-331.

On the Significance of Tourism Website Evaluations

Magda Antonioli Corigliano[a],
Rodolfo Baggio[a, b]

[a] Master in Economics and Tourism
Bocconi University, Milan, Italy
{magda.antonioli, rodolfo.baggio}@unibocconi.it

[b] School of Tourism and Leisure Management
The University of Queensland, Australia

Abstract

Website evaluation methods are an important tool to gather information for the development and the management of a website in order to ensure a good acceptance by the users. Mainly based on detailed questionnaires administered to actual or potential users, this activity can be costly and time consuming. This paper present methods, based on the statistical bootstrapping techniques, to derive confidence intervals for the evaluations produced by a certain sample of users. It is shown that relatively small (less than 10%) confidence intervals can be achieved even with small sample sizes (less than 20).

Keywords: website evaluation; confidence intervals; sampling size; bootstrap methods.

1 Introduction

The World Wide Web has become a ubiquitous tool to find information and conduct business, and it is still growing at a very high rate. As a primary means to disseminate information to the public, this environment requires ongoing performance measurements (such as number of visitors or online sales) concerning the extent to which their websites are successfully presenting and conveying information and services the public needs to access and use. Achieving good results is directly connected to the quality level of the implementations (see for example Morrison et al., 2004).

The Internet age has allowed the development of new ways for producing and distributing tourism services. With this, the tourism industry faces a whole new series of challenges and opportunities (Buhalis, 2003). Web-based approaches and technologies are helping tourism suppliers and agencies reduce service costs and attract customers. A website looks to be a major (and, probably, it will be the only one in the future) tool to conduct business in the tourism field. According to PhoCusWright's (2005) estimates, for example, that in the U.S. online sales will be 35% of the travel market in 2005 and more than 50% in 2006.

1.1 Tourism websites evaluations

Basically, a website is a software application directed to a vast and typically not well skilled population of users. As such, all the considerations and the issues regarding its *usability* are extremely important for assuring its acceptance by the users. For the implications and the effects it may have on the success of an organization, this communication channel faces the big challenge of trying to attain the most favourable reception by the visitors.

The relationships between the satisfaction in browsing a website and the brand image (both *virtual* and *real*) or the business results have been recognized in a number of studies (Baggio, 2005; Hahn & Kauffman, 2002; Rajgopal et al., 1999). The evaluation of the quality characteristics of a website has assumed, therefore, a fundamental importance.

The evaluation of a website, whether as part of a formal planning and management process or as a stand-alone activity, is important for at least two reasons: it can provide managers with key information useful to maximize the returns (tangible or intangible) a realization can provide, and it can help studying the behaviour of the users and their reactions to the contents and services offered online.

This importance has been well acknowledged by academics, consultants and practitioners, so that a great number of studies have been published in the last years. In what follows our main reference will be the list provided by the University of Trento eTourism group (UNITN, 2005). It contains 249 titles of scientific works on the general subject of web evaluation, 50 titles of publications specifically dedicated to the tourism industry and 60 addresses of websites dedicated to the topic.

There is no universally accepted method or technique for a website evaluation. While browsing the papers of UNITN list, the feeling is that 100 authors mean 100 different schemes. Nevertheless, the evaluation methods can be grouped in two broad categories:

- *automated methods*: based on automatic tools able to capture, mainly, technical characteristics such as response times, conformance to language standards, or structural coherence.

- *heuristic usability methods*: where casual or expert users judge whether each element of a web interface follows pre-determined usability and aesthetic principles.

The second one is by far the most used class. No more than 5% of UNITN list works discuss automatic methods, and, in many cases, only to perform partial assessments in a heuristic framework.

Even if, as said, there are no commonly agreed practices, these methods, in essence, follow a rather common general path:

- the investigator establishes a list of characteristics, typically grouped in classes such as: graphical quality, textual contents, interactive services, technical attributes etc.;
- the list (ranging from a few elements to some hundreds) is transformed into a questionnaire;
- the questionnaire is administered to a number of users that are asked to inspect the website (or websites) evaluating the questionnaire items by assigning a mark;
- the final score is derived from the single item evaluations (usually by averaging).

The number of evaluators (i.e., the sample size) is an important element to derive a *significant* result from these assessments. Classical statistical procedures have well grounded methodologies for estimating the ideal size of a sample depending on the population parameters. The range is typically of the order of magnitude of 10^2 - 10^3 (see for example Cochran, 1977).

Evaluations such as those described above may be quite costly and time consuming, mainly if we agree that, to be an effective way of managing a website, they should be repeated at regular intervals. This may be a strong deterrent for many organizations, and may prevent them from performing this activity with possible detrimental effects on the effectiveness of their investments and their satisfaction in using the Internet channels. A possible way of addressing this issue is the reduction of the number of evaluations to be performed, without affecting the significance of the outcomes.

Aim of this paper is to discuss this topic and to give some practical guidance on it. Apart from the obvious theoretical interest, the subject has also a strong practical value, it is therefore quite important to be able to determine the correct and minimum requirements for a significant assessment.

The remainder of this paper is organized as follows. The next Section presents the background, mainly statistical, to the problem and the methodology used for this work. Section 3 contains and discusses the results obtained. Finally, we present our conclusive remarks in Section 4.

Readers not interested in the discussion on methodological issues can skip the related sections (2 and 3) and examine the conclusive remarks (section 4).

2 Background and Methods

As said in Section 1, the discussion on the sampling size and its validity, is not very frequently found in the papers dealing with website evaluations (whether tourism or not). Moreover, a vast variety of sizes is used; an inspection of 130 papers taken from the UNITN list (the list is available from the authors, it is not fully quoted here for space reasons) give the situation depicted in Fig. 1. More examples and a summary table can be found in the review by Morrison et al. (2004).

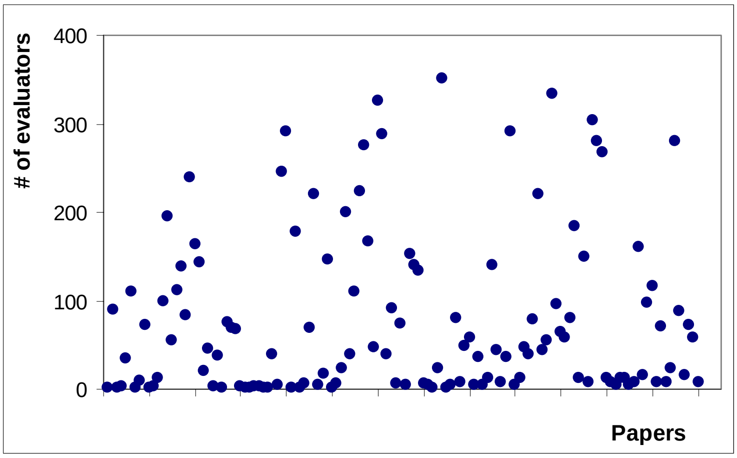

Fig. 1. Number of evaluators used in 130 website evaluation studies

Almost everything, from a single person to a "classical" sample of some hundreds is used. It must be noted that a noticeable number of papers gives no indications at all on the size of the users' sample used; we may suppose that these works are based on the authors (*expert*) judgement of the cases presented. While this is acceptable if the aim of the inspection is solely to assess the presence of certain features, any form of "judgement" on these should be backed by correct statistical treatment of the sample of observations.

It can be argued that in seeking effects, rather than looking for the quantification of the characteristics of a population, the sample size can be less important (Anderson & Vingrys, 2001). However, an estimate of the significance of the results obtained is greatly valuable.

We may assume that, implicitly or explicitly, the main field in which evaluation activities are performed is the one of software usability. In this domain, the discussion on the ideal evaluators sampling size counts a number of works (e.g., Nielsen, 1994; Nielsen & Molich, 1990; Virzi, 1992). Their conclusion is, generally, that a limited number of users is sufficient to give, with good reliability, meaningful results, even if there are different interpretations on the actual size of the sample (Faulkner, 2003; Woolrych & Cockton 2001, Gray & Salzman, 1998a, 1998b).

The experience of the evaluators has also been invoked as a factor that may reduce the number of evaluators needed (Karoulis & Pombortsis, 2004).

It is questionable, though, whether these conclusions can be applied to a website evaluation. Here the question is not the one of finding most of the possible "usability problems", basically of technical nature. It is rather the one of trying to establish a general satisfaction level expressed by a population of users consulting a website. In the tourism field this satisfaction can be one of the main determinants of a travel decision or of the choice of a service provider (Fesenmaier et al., 2003; Jeng & Fesenmaier, 2002; Holland & Menzel Baker, 2001).

The issue to be addressed is the reliability of a small (the smallest possible) sample of users needed to obtain a reasonably significant assessment of the quality of a website.

The problem of using small samples is well known in many disciplines, mainly in those where the replicability of the measurements is problematic (astronomy and astrophysics, for example). In these cases, several methods to derive confidence levels for observations with a very limited number of events have been formulated (Gehrels, 1986; Regener, 1951).

One different possibility to assess the reliability of a given (small) sample is to use newer statistical techniques such as the bootstrap methods introduced by Bradley Efron (Efron, 1979; Efron & Tibshirani, 1993).

Bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample. The method (Efron & Tibshirani 1993) allows a researcher to obtain an approximation of the distribution of a statistical estimator, in the absence of a priori information about the true distribution of the estimator or of the original data.

From a given sample, a large number (usually 1500-2000) of new data sets is generated by drawing, with replacement, a certain number of observations from the original sample. The estimator is calculated for each new data set. The resulting empirical distribution of estimator values is used to approximate its true distribution.

Given a set of independent observations x_1, x_2, … x_n, a parameter that can be defined as some function $\theta = G(x)$ of the values in the population and a statistic that is the same function of the observations $\theta' = G(x)$, the bootstrap estimates the sampling distribution $F_\theta(x)$ of that function. Bootstrap samples are repeatedly drawn from the estimated population. The function (e.g. the mean) is evaluated for each bootstrap sample, giving a set of bootstrap values θ'_{B1}, θ'_{B2}, …, θ'_{Bn}. The empirical distribution of these bootstrap values $F'_B(x)$ estimates the theoretical sampling distribution $F'_\theta(x)$. The bootstrap distribution $F'_B(x)$ is also used to estimate standard errors or to construct a confidence interval for the statistic of interest.

Standard parametric confidence intervals can provide a measure of significance for a statistical estimator. They require, however, the acceptance of normality assumptions and a large sample size for their validity. With the bootstrap method, it is possible to calculate the confidence interval of the parameter estimated from the distribution generated by the replications, without being forced to accept normality hypotheses. Usually, the percentile confidence interval method is used. It uses the $\alpha/2$ and $1- \alpha/2$ quantiles of $F'_B(x)$ as $1-\alpha$ confidence interval for the parameter.

The bootstrap methods have seen many applications in several fields and have proved useful means for sample size determination based on pilot experiments in the preparation of large surveys (Mak, 2004).

3 Results and discussion

For this study, a set of 250 evaluations of a single tourism destination website has been collected. The evaluations follow a scheme already used in previous works (Antonioli & Baggio, 2002; 2004). A predetermined list of items is provided for inspection and evaluation. The list comprises elements such as the quality of graphical representations, the balance between text and pictures, the clarity of the descriptions, etc. The evaluation is qualitative, the visitors express their appreciation of the various website features by means of a score (1=min to 5= max).

A plot of the progressive average for the 250 evaluations (Fig. 2) shows that, after a certain number (limited) of evaluations, the average looks stabilizing in a small interval (±5%) around the "final" average.

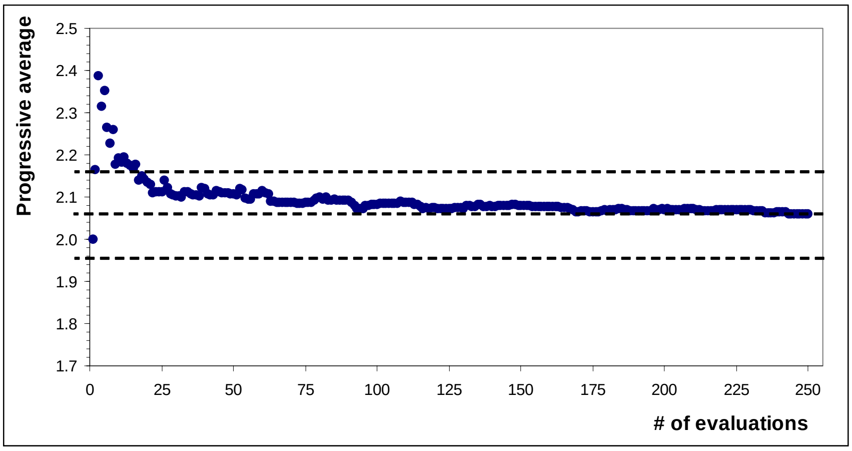

Fig. 2. Progressive average for the evaluations. Dotted lines represent the interval ±5% around the whole sample average (2.06)

This effect, the point at which the averages "stabilize", depends, rather obviously, on the arrangement of the data. However, it seems to occur for a rather low number of observations. Fig. 3 shows the results obtained in 100 realizations of a random re-ordering of the values.

If the original sample is considered to be the population, what sample (subsample) of these data is needed to obtain an average whose margin of error (at a predetermined confidence level) is lower than a certain value? In other words: if a subsample of, say, 15 random evaluations is taken, what is the margin of error (at a certain confidence level) of considering this one the "real evaluation" of the website under investigation? The bootstrap methods can provide an answer to these questions: the estimate of the variability of the observations.

In the present study, the question has been addressed by applying the bootstrap method to a series of random subsamples drawn from the original set of evaluations. The statistical estimator used is the arithmetic mean. Nonparametric bootstrap percentile confidence intervals are used to infer the observed significance level of the effects. Following the considerations by Andrews and Buchinsky (2000) on the most reliable choice of bootstrap repetitions, the number used is 1000.

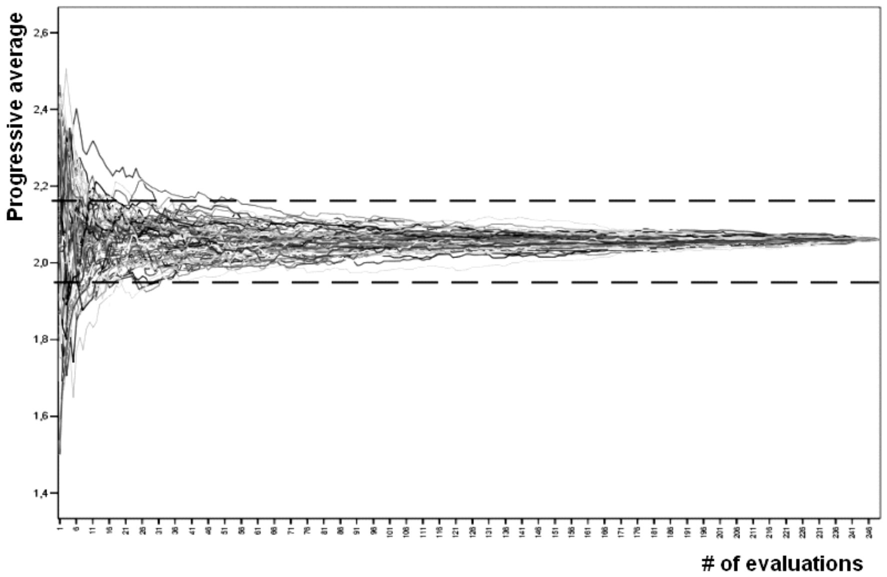

Fig. 3. Progressive average for the evaluations. 100 random re-orderings of the original data. Dotted lines represent the interval ±5% around the whole sample average (2.06)

Ten series for each of 9 subsamples (10, 15, 20, 25, 50, 75, 100, 125 and 150 elements) have been randomly chosen from the whole set of values. Each series has been “bootstrapped” and the arithmetic mean has been calculated. The bootstrap distribution of each value was compiled, and the 5th and 95th percentiles of the empirical distribution formed the limits for the 95% bootstrap percentile confidence interval.

The standard bootstrap confidence interval is accurate only for statistics with an approximately symmetric (normal) sampling distribution (Shao & Tu, 1995). Symmetry of the distributions has been verified by testing the skewness coefficient for all the distributions obtained. All calculations were performed by means of procedures developed for Matlab® (version 6.5, release 13, 2002; The MathWorks Inc.: Natick, MA).

The values obtained are shown in Table 1 and Fig. 4. They are the average over the ten realizations for each subsample. The last row contains, for comparison, the results for the whole 250 items set of observations (see also Fig. 5).

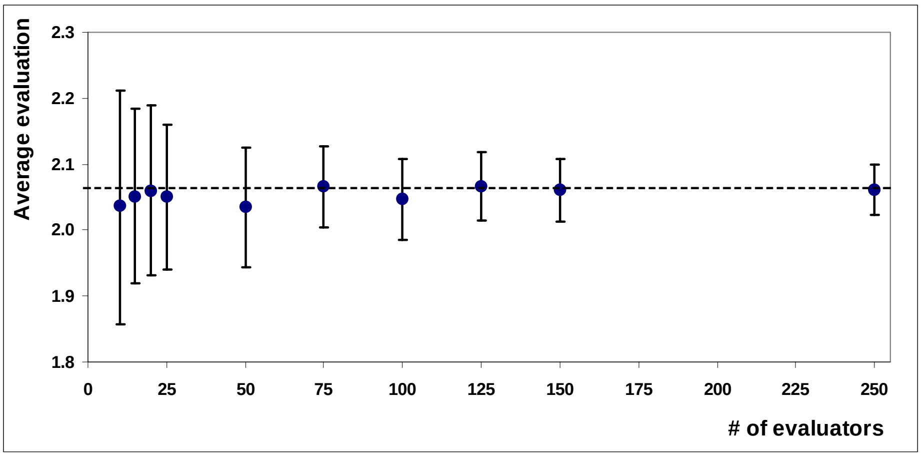

Fig. 4. Average scores calculated for different evaluators sample sizes. The error bars represent the confidence intervals, dotted line is the average calculated for the whole original set of evaluations.

Table 1. Confidence intervals (at 95%) for the average scores calculated for different sample sizes. Values for whole original set of evaluations are added for comparison.

# of evaluators	Ave. evaluation	95% Conf. Interval
10	2.04	8.7%
15	2.05	6.5%
20	2.09	6.3%
25	2.05	5.4%
50	2.04	4.4%
75	2.07	3.0%
100	2.05	3.0%
125	2.07	2.5%
150	2.06	2.3%
Whole sample	***2.06***	***1.8%***

As it can be seen, the confidence intervals are relatively small (less than 10%), even with small numbers of evaluators (15 or 20). The results can be used to give guidance in deciding what kind of "precision" is to be attained in conducting an evaluation. For example, we may say that, at 95% CL, there is a 6.5% error on the final evaluation estimate if it is conducted by 15 people (randomly chosen among the users).

4 Conclusive Remarks

Assessing the quality of a website as perceived by the users is an important process for any organization using the Internet as a communication, promotion or business medium. This is even more important for a tourism organization given the significance the Web has for this sector.

The typical evaluation of a website is performed by using a heuristic usability test. This is done, basically, by selecting a sample of users and administering a questionnaire asking for the assessment of a number of items. Time and cost considerations, especially important for small and medium organizations, the vast majority in the tourism sector, suggest to find ways to reduce the number of evaluators needed while assuring a reasonable statistical significance of the results.

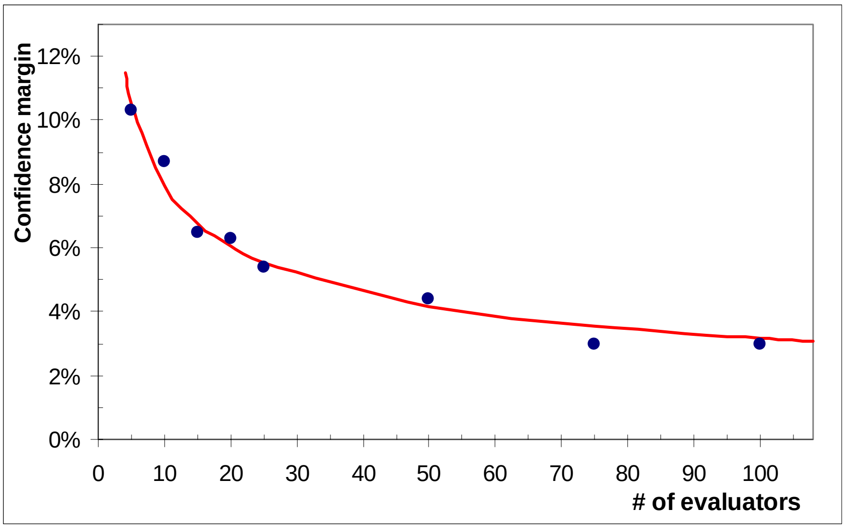

Fig. 5. Confidence margins as function of the number of evaluators for tourism websites assessments (for sample sizes ≤ 100, solid line is the best fit of the data)

A statistical analysis technique, the bootstrap, can prove useful in giving such an answer; by applying it, this study has shown that, even with relatively small sample sizes (15-20 evaluators) it is possible to obtain reasonable confidence intervals (less than 10%).

The results can be used to either assess the confidence margins (CM) of a given sample of evaluators (N_E) or to plan in advance the website evaluation according to the objectives an organization has. For example, a larger sample (and smaller CMs) can be needed in an initial phase of website development, while smaller samples can be used to perform periodic evaluations in absence of major modifications to the website. Fig. 5 can be used as a quick reference. The line represents the best fit in a partial region (sample size ≤ 100).

For the math-savy the formula is: $CM = 0.26\ N_E^{-0.48}$ with $R^2 = 0.98$.

The obvious limitation of this study lies in the fact that a single case has been used. However, aim of this work was more to present a technique to validate the results of a tourism website evaluation than to provide conclusive answers.

The methods proposed in this paper have the advantage of being relatively easy and fast to implement. In this way, hopefully, tourism organizations can be better inclined at using user evaluations of their websites in order to improve their quality and acceptance by the users, so important for the development of their business.

References

Anderson, A. J., & Vingrys, A. J. (2001). Small Samples: Does Size Matter? *Investigative Ophthalmology & Visual Science*, 42(7), 1411-1413.

Andrews, D. W. K., & Buchinsky, M. (2000). A Three-Step Method for Choosing the Number of Bootstrap Repetitions. *Econometrica*, 68, 23-51.

Antonioli Corigliano, M., & Baggio, R. (2004). Italian Tourism on the Internet - New Business Models. In Weiermair, K., Mathies, C. (Eds.), *The Tourism and Leisure Industry - Shaping the Future* (pp. 301-316). New York: The Haworth Press.

Antonioli Corigliano, M., & Baggio, R. (Eds.) (2002). *Internet e Turismo.* Milano: Egea.

Baggio, R. (2005). The Relationship Between Virtual and Real Image of Tourism Operators. *eReview of Tourism Research*, 3 (5), http://ertr.tamu.edu.

Buhalis, D. (2003). *eTourism: Information technology for strategic tourism management*. Harlow, UK: Pearson/Prentice Hall.

Cochran, W. G. (1977). *Sampling Techniques (3rd ed.)*. New York : John Wiley & Sons.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments and Computers*, 35(3), 379-383.

Fesenmaier, D. R., Gretzel, U., Hwang, Y., & Y. Wang (2003). The future of destination marketing: e-Commerce in travel and tourism. *International Journal of Tourism Sciences*, 3 (2), 191-200.

Gehrels, N. (1986). Confidence limits for small numbers of events in astrophysical data. *Astrophysical Journal*, 303, 336-346.

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

Gray, W. D., & Salzman, M. C. (1998b). Repairing damaged merchandise: A rejoinder. *Human-Computer Interaction*, 13(3), 325-335.

Hahn, J., & Kauffman, R. J. (2002). Measuring and Comparing the Effectiveness of E-Commerce Website Design", 14th Workshop on Information Systems and Economics (WISE), IESE Business School, University of Navarra, Barcelona, Spain, Dec 14-15. (Working Paper: http://www.jungpil.net/research.html, last access: September 2005).

Holland, J., & Menzel Baker, S. (2001). Creating Site Brand Loyalty. *Journal of Interactive Marketing*, 5 (4), 34-45.

Jeng J., & Fesenmaier, D.R. (2002). Conceptualizing the Travel Decision-Making Hierarchy: A Review of Recent Developments. *Tourism Analysis*, 7 (1), 15-32.

Karoulis, A., & Pombortsis, A. (2004). The Heuristic Evaluation of Web-Sites Concerning the Evaluators' Expertise and the Appropriate Criteria List. *Informatics in Education*, 3 (1), 55-74.

Mak, T. K. (2004). Estimating variances for all sample sizes by the bootstrap. *Computational Statistics & Data Analysis*, 46, 459-467.

Matlab® (2002), Version 6.5, Release 13, June 2002, Natick, MA: The MathWorks Inc.

Morrison, A. M., Taylor, J. S., & Douglas, A. (2004). Website Evaluation in Tourism and Hospitality: The Art Is Not Yet Stated. Journal of Travel &Tourism Marketing, 17 (2/3), 233-251.

Nielsen, J. (1994). Heuristic Evaluation. In Nielsen & Mack (Eds.), *Usability Inspection Methods(pp.* 25-62). New York: John Wiley & Sons.

Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. *Proceedings of CHI Conference on Human Factors in Computing Systems* (pp. 249-256). New York: ACM.

PhoCusWright (2005). *Online Travel Overview: Market Size and Forecasts 2004-2006*. Sherman, CT: PhoCusWright.

Rajgopal, S., Venkatachalam, M., & Kotha, S. (2001). *Does the Quality of Online Customer Experience Create a Sustainable Competitive Advantage for E-commerce Firms?* GSB research paper series #1666, Stanford University Graduate School of Business.

Regener, V. H. (1951). Statistical Significance of Small Samples of Cosmic Ray Counts. *Physical Review*, 84, 161-162.

Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap.* New York: Springer-Verlag.

UNITN (2005). *Extended bibliography related to Website Quality Evaluation* and *Bibliography related to Tourism Website Quality Evaluation*. Last updated: 1 September 2005. Online: http://www.economia.unitn.it/etourism/risorseQualita.asp, last access: September 2005.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34(4), 457-468.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), *Proceedings of IHM-HCI 2001 Conference* (Vol. 2, pp.105-108). Toulouse (F): Cépadèus.

Acknowledgements

A large part of the studies on website assessment use students as evaluators. The authors wish to express their grateful thanks to all the students, from any country or school, that have contributed so strongly to the increase of the knowledge in this particular area.