Published in: *Software Quality and Software in Internet Time*, Springer Publishing

Low Cost Usability Testing

Drs. Erik P.W.M. van Veenendaal CISA
Improve Quality Services BV / Eindhoven University of Technology

Usability is an important aspect of software products. However, in practice not much attention is given to this issue during testing. Testers often do not have the knowledge, instruments and/or time available to test for usability. This paper identifies the Heuristic Evaluation and the Software Usability Measurement Inventory (SUMI) testing techniques as a possible solution to these problems. The focus of this paper will be on the latter. Heuristic evaluation involves having a small set of evaluators examine the interface and judge its compliance with recognized usability principles)"the heurtistics"). SUMI is a rigorously tested and validated questionnaire based method to measure software quality from a user's perspective. Using SUMI the usability of a software product or prototype can be evaluated in a consistent and objective manner[1]. The technique is supported by an extensive reference database and embedded in an effective analysis and reporting tool.SUMI has been applied to a great number of projects. This paper discusses three practical applications. Results, usability improvements, cost and benefits are described.

1. A Closer Look at Usability

Several studies have shown that in addition to functionality and reliability, usability is a very important success factor (Nielsen,1993) (MultiSpace,1997). But although it is sometimes possible to test the software extensively in a usability lab environment, in most situations a usability test has to be carried out with minimum resources.

Usability of a product can be tested from mainly two different perspectives: "ease-of-use" and "quality-in-use". Quite often the scope is limited to the first perspective. The ease or comfort during usage is mainly determined by characteristics of the software product itself, such as the user interface. Within this type of scope usability is part of the product quality characteristics. The usability definition of ISO 9126 is an example of this type of perspective:

Usability the capability of the software to be understood, learned, used and liked by the user, when used under specified conditions (ISO 9126-1,2001)

Two techniques that can be carried out at reasonable costs to evaluate the usability product quality are heuristic evaluation and checklists. These techniques have the disadvantage that the real stakeholder, i.e. the user, often isn't involved.

In a broader scope usability is determined by using the product in its (operational) environment. The type of users, the tasks to be carried out and the physical and social aspects that can be related to the usage of the software products are taken into account. Usability is defined as "quality-in-use". The usability definition of ISO 9241 is an example of this type of perspective:

[1] Research has shown that the SUMI scores are at least 80% reliable (Kelly, 1994).

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

Usability the extent to which a product can be used by specified users to achieve goals with effectiveness, efficiency and satisfaction in a specified context of use (ISO 9241-11,1996)

Clearly these two perspectives of usability are not independent. Achieving "quality-in-use" is dependent on meeting criteria for product quality. The interrelationship is shown in Figure 1.

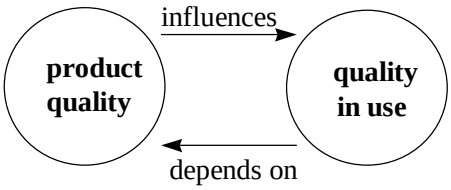

Figure 1 : Relationship between different types of usability

Establishing test scenarios, for instance based on use cases (Jacobson,1992), can be applied to test usability in accordance with ISO 9241. However, usability testing with specified test cases / scenarios is a big step for most organisations and often not even necessary. From a situation where usability is not tested at all easier techniques are needed, preferably ones that involves users, are reliable, but only require limited resources. Of course these techniques do not have the same level of coverage as a full usability test (Rubin,1994), however in most situations they suffice.

2. Heuristic Evaluation

Heuristic Evaluation is a systematic examination (inspection) of a finished product, design or prototype from the point of view of its usability by intended end users. The process is by preference a team effort that includes developers, end-users, application or domain experts, and usability specialists in collaboration (Nielsen en Mack, 1994). A major benefit of the method is its own usability: it is easy to learn. Experienced inspection teams have been known to detect 100 usability faults per hour. The collaborative nature of these inspections means that users and developers understand the relationship between the user interactions and the design constraints and decisions. These inspections can be performed at any stage of development, from the assessment of prototypes to the finished products. but, of course, the cost of fault correction increases, the later the inspections take place.

Heuristic evaluation is performed by having each individual evaluator inspect the interface alone. Only after the evaluations have been completed are the evaluators allowed to communicate and have their findings aggregated. This procedure is important to ensure independent and unbiased evaluations from each evaluator.

The inspection are guided by a set of rules or heuristics for good user interface design. These are used as a framework for identifying and categorising usability faults. Nielsen promotes a (popular) set of ten design heuristics (Nielsen,1993) summarized in table 1. His books, Usability Engineering and Designing Web Usability are packed with usability design guidelines from which you could extract your own preferred rules if you cared to.

1.	Visibility of system status
2.	Match between system and the real world
3.	User control and freedom
4.	Consistency and standards
5.	Error prevention

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

6.	Recognition rather than recall
7.	Flexibility and efficiency of use
8.	Aesthetic and minimalist design
9.	Help users recognize, diagnose, and recover from errors
10.	Help and documentation

Table 1: Usability heuristics

Other rules can be used, and several collections of checklists are available. Other books and web sites provide substantial checklists of usability design issues that could be used but probably need to be summarised to be useful as inspection rules. When you test, failure to adhere to these rules will be raised as incidents.

2. What is SUMI?

Within the European ESPRIT project MUSiC [ESPRIT 5429] a method has been developed that serves to determine the quality of a software product from a user's perspective. Software Usability Measurement Inventory (SUMI) is a questionnaire-based method that has been designed for cost-effective usage. Software Usability Measurement Inventory (SUMI) is a solution to the recurring problem of measuring the user's perception of the usability of software. It provides a valid and reliable method for the comparison of (competing) products and differing versions of the same product, as well as providing diagnostic information for future developments (Kirakowski and Corbett, 1993). As stated before SUMI (and the Heuritic Evaluation) are not a substitution of a full usability test. In fact, SUMI "only" measures the user's perception of usability – one aspect of usability testing.

SUMI consists of a 50-item questionnaire devised in accordance with psychometric practice. Each of the statements is rated with "agree", "undecided" or "disagree". The following sample shows the kind of questions that are asked:

- This software responds too slowly to inputs.
- I would recommend this software to my colleagues.
- The instructions and prompts are helpful.
- I sometimes wonder if I am using the right command.
- Working with this software is satisfactory.
- The way that system information is presented is clear and understandable.
- I think this software is consistent.

The SUMI questionnaire is available in English (UK and US), French, German, Dutch, Spanish, Italian, Greek and Swedish.

SUMI is intended to be administered to a sample of users who have had some experience in using the software to be evaluated. In order to use SUMI reliably, a minimum of ten users is recommended based on statistical theory. Based on the answers given and statistical concepts, usability scores are calculated. Of course SUMI needs a working version of the software before SUMI can be performed. This working version can also be a prototype or a test release.

One of the most important aspects of SUMI has been the development of the standardisation database, which now consists of usability profiles of over 2000 different kinds of applications. Basically any kind of application can be evaluated using SUMI as long as it has user input through keyboard or pointing device, display on screen and some input and output between secondary memory and peripheral devices. When evaluating a product or series of products using SUMI, one may either do a product-against-

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

product comparison or compare each product against the standardisation database, to see how the product that is being rated compares against an average state-of-the-market profile.

SUMI gives a global usability figure and additional readings on five sub-scales:

- *Efficiency*: degree to which the user can achieve the goals of his interaction with the product in a direct and timely manner
- *Affect*: how much the product captures the user's emotional responses
- *Helpfulness*: extent to which the product seems to assist the user
- *Control*: degree to which the user feels that he, and not the product, is setting the pace
- *Learnability*: ease with which a user can get started and learn new features of the product.

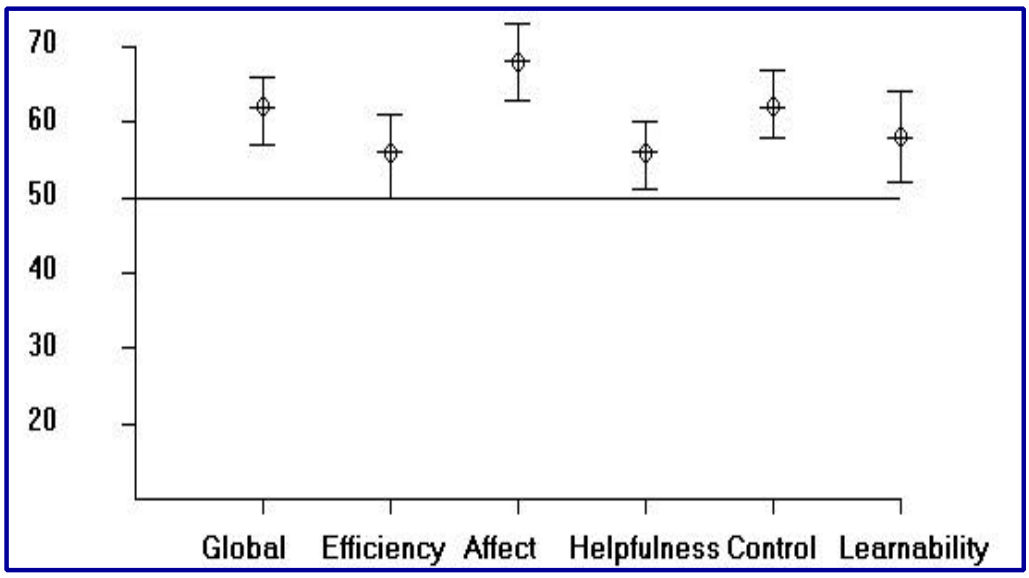

Figure 2: a sample profile showing SUMI scales

Figure 2 shows an example of SUMI output; it shows the scores of a test and the spreading of these scores (measured by the standard deviation) against the average score of the reference database, reflected by the value 50. Consequently the usability scores shown in the sample profile are positive, i.e. more than state-of-the-art, with a reasonable level of spreading.

SUMI is the only available questionnaire for the assessment of software usability, which has been developed, validated and standardised on a European-wide basis. The SUMI sub-scales are referenced in international ISO standards on usability (ISO 9241-10,1994) and software product quality (ISO 9126-2,1997). Product evaluation with SUMI provides a clear and objective measurement of the user's view of the suitability of software for his tasks.

This provides a solid basis for specialised versions of SUMI. Recently MUMMS has been developed for multimedia products (Measuring Usability of Multi Media Systems).

Any SUMI test must be carried out by asking people to perform realistic, representative tasks. Employing a method such as usability context analysis (NPL,1995) helps identify and specify in a systematic way the characteristics of potential users, the tasks to be carried out and the circumstances of use. The results can be used to identify the various user groups and to define how these user groups are to be represented during testing.

3. Practical Applications

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

3.1 Project 1: Project Management Package

3.1.1 Approach

A software package offering project administration and control functionality was subjected to the usability evaluation by means of SUMI. The software package is positioned as a multi-project system for controlling the project time, e.g. in terms of scheduling and tracking, and managing the productivity of projects, e.g. in terms of effort and deliverables. The package has been developed by a Dutch software house that specialises in the development of standard-software packages.

The SUMI test was part of an acceptance test carried out on behalf of a potential customer. Usability was an important characteristic due to the very high number of users, different user groups, their inexperience with project management software and the great variety of information needs. It was even looked upon as the critical success factor during implementation. Two main user group were distinguished. One user group was mainly involved in the input of effort and time spent. Operability and efficiency are of especially great importance for this user group. Another user group was characterised as output users. Especially receiving the right management information is important for the output users. A SUMI test was carried out for each user group.

Specific acceptance criterion were applied regarding the usage of the SUMI technique for the usability evaluation. SUMI provides quantitative values relating to a number of characteristics that lead to a better understanding of usability. As part of the acceptance test, the SUMI scale was used that provides an overall judgement of usability, the so-called “global scale”. Based on the data in the SUMI database, it can be stated that the global score has an average value of 50 in a normal distribution. This means that, by definition, for a value exceeding 50 the user satisfaction is higher than average. For the test of the project management package the acceptance criteria required that for each user group the global scale and the lower limit of the 95% confidence interval must both exceed the value of 50.

3.1.2 Results

The "global scale" regarding both user groups was below the desired the value of 50. For the input user group the score was even a mere 33. The output user group showed a slightly better score. Not only the “global scale” but also most other sub-scales were scoring below 50.

Because the results did not meet the acceptance criteria a number of usability improvement measures needed to be taken. Examples of measures that were taken based on the results of the SUMI test are:

- extension and adaptation of the user training
- optimisation of efficiency for important input functions
- implementation of specific report generation tools for the output user with a clear and understandable user-interface.

3.2 Project 2: PDM system

3.2.1 Approach

A Product Data Management System (PDMS) was implemented in the R&D department of a large copier manufacturer. During the trial phase usability appeared to be an issue and could have become a

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

major risk factor during implementation. The time and effort needed to be spent on usability formed a point of discussion between development and the user organisation. It was decided to apply SUMI to acquire an insight into the current user perception of the PDMS.

A number of randomly selected users who were involved in the PDMS trail phase were requested to fill out the questionnaire. Twenty six users were selected, of whom twenty-one returned the questionnaire. Six users stated that they didn't use the PDMS often enough. The feedback thus resulted in a 77% response rate.

3.2.2 Results

The table below shows the overall scores for the various SUMI sub-scales:

	Global	Efficiency	Affect	Helpfulness	Control	Learnability
Median	36	31	43	36	36	35

Table 1: SUMI scores PDMS

The various scores were relatively low all round. There didn't seem to be too large of a divergence of opinion, except perhaps for learnability. An analysis of the individual user scores did not show any real outlier (see next table). Two users (one and five) had an outlier score for one scale (too high). Since it was only on one scale, they were not deleted from the respondent database.

	G	E	A	H	C	L
User 1	60	52	59	**69**	47	32
User 2	57	48	53	62	41	61
User 3	25	19	46	35	22	33
User 4	17	14	28	11	26	23
User 5	61	63	55	44	**60**	64
User 6	24	23	23	36	22	14
User 7	53	62	44	..	..	..
User ..	..	..	..	..	..	..

Table 2: SUMI scores per user

As stated earlier the various scores were relatively low all round. In general one can say that the user satisfaction regarding the system is too low and corrective action is needed. Some more detailed conclusion were:

- *Efficiency*
  According to the users PDMS doesn't support the user tasks in an efficient manner. Too many and too difficult steps must be performed. Consequently, one cannot work efficiently and feels that the system is insufficiently customised to users' needs.
- *Helpfulness*
  An important conclusion is the fact that the messages are often not clear and understandable; as a consequence the system doesn't provide much help in solving a problem. The possibilities provided to the user in each situation are not clearly shown.
- *Control*
  Users often have the feeling of not being control and find it difficult to let the system behave in the way they want it to. They feel safe when they only use commands they know. However, they do find it easy to jump from one task to another.

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

On the basis of the SUMI evaluation it was decided to define a number of follow-up actions:

- a detailed analysis of the problems as perceived by the users. A number of users are interviewed and asked to explain, by means of practical examples, the answers given to the SUMI questions;
- a study on outstanding change requests and probably an increase in their priority;
- an improved information service to the users on changed functionality to provide them with more knowledge on how to operate the system;
- a re-evaluation of the training material with user representatives;
- a SUMI test was to be carried out on a regular basis (every two/three months) to track the user satisfaction during implementation of the PDMS.

Currently the follow-up is in progress and no new SUMI test has yet taken place. Consequently, nothing can be said regarding the improvement of the usability. However, by means of the SUMI test, usability has become a topic within the PDMS project that gets the attention (time and effort) it apparently needs.

3.3 Project 3: Intranet site

3.3.1 Approach

By means of MUMMS, the specialised multimedia version of SUMI, the usability of an intranet site prototype of a large bank was evaluated. The intranet site was set up by the test services department to become better known and to present themselves to potential customers. Since during the test only a prototype version of the intranet site was available some pages were not yet accessible. A special sub-scale has been introduced for MUMMS, with the objective of measuring the users' multimedia "feeling":

- *Excitement*: extent to which end-users feel that they are "drawn into" the world of the multimedia application.

In total, ten users (testers) were involved in the MUMMS evaluation. The set of users can be characterised in the following ways:

- not involved during the development of the intranet site
- potential customers
- four users with Internet experience
- six users without Internet experience
- varying age and background (job title).

3.3.2 Results

The table below shows the overall scores for the various MUMMS sub-scales:

	Affect	Control	Efficiency	Helpfulness	Learnability	Excitement
average score	69	74	62	67	67	68
median	71	77	67	69	67	72
standard deviation	9	12	11	8	6	12

Table 3: Overall MUMMS score table

The various scores were moderately high. However, there seems to be a divergence of opinion on the

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

control and excitement scales. Some low scores pull down the control and efficiency scales (see next table). Two users from the sample gave exceptionally low average scores. They were analysed in detail but no explanation was found.

	A	C	E	H	L	E	Average
User 1	71	81	67	71	74	77	73
User 2	74	74	74	71	67	71	72
User 3	81	84	67	67	74	74	74
User 4	54	51	54	57	64	44	**54**
User 5	71	74	43	58	55	76	63
User 6	64	84	67	81	67	69	72
User 7	51	81	74	54	74	64	66
User 8	71	81	64	74	71	81	73
User 9	77	81	76	84	77	74	78
User 10	64	47	51	57	57	44	**53**

Table 4: MUMMS scores per user

As stated the usability of the intranet site was rated moderately high from the users' perspective, although there seemed to be a lot of divergence in the various user opinions. Some more detailed conclusions were:

- *Attractiveness*
  The attractiveness score is high (almost 70%). However some users (4, 7 and 10) have a relatively low score. Especially the questions "this MM system is entertaining and fun to use" and "using this MM system is exiting" are answered in different ways. It seems some additional MM features should be added to further improve the attractiveness for all users.
- *Control*
  A very high score for control in general. Again two users can be identified as outliers (4 and 10) scoring only around 50%, the other scores are around 80%. Problems, if any, in this area could be traced back to the structure of the site.
- *Efficiency*
  The average score on efficiency is the lowest, although still above average. Users need more time than expected to carry out their task, e.g. to find the right information.

On the basis of the MUMMS evaluation it was decided to improve the structure of the intranet site and to add a number of features before releasing the site to the users. Currently an update of the intranet site is being carried out. A MUMMS re-evaluation has been planned to quantify the impact of the improvement regarding usability.

4. Applicability of SUMI

On the basis of tests carried out in practice, a number of conclusions can be drawn regarding the applicability of SUMI and MUMMS:

- It is easy to use and involves not many costs. This applies both to the evaluator and the customer. On average a SUMI test can be carried in approximately three days; this includes the time needed for a limited context analysis and reporting.
- During testing the emphasis is on finding defects, this often results in only negative quality indications, e.g the number of defects found. SUMI however, provides an objective opinion that can also be a positive quality indicator, e.g. a SUMI score of 70 or more.
- The usability score is split into various aspects, making a thorough, more-detailed evaluation

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

possible (using the various output data).
- SUMI provides, after detailed analysis and discussion, directions for improvement and directions for further investigation. SUMI can also be used as a risk analysis method to determine whether a more detailed usability testing is necessary.

However, some disadvantages can also be mentioed:
- A running version of the system must be available; this implies SUMI can only be carried out at a relatively late stage of the project.
- A high number of users (minimum of ten) with the same background are needed to fill out the questionnaire. Quite often the implementation or test doesn't involve ten or more users belonging to the same user group.
- The accuracy and level of detail of the findings is limited (this can partly be solved by adding a small number of open questions to the SUMI questionnaire). In practice a SUMI evaluation is often carried out in co-operation with a Heuristic Evaluation, the latter can in such a case provide a thorough interpretation of the SUMI score and concrete direction for improvement.

5. Conclusions

A system's end users are *the* experts in using the system to achieve their goals. Therefore, their voices should be listened to when the system is being evaluated. SUMI does precisely that: it allows quantification of the end users' experience with the software and it encourages the tester to focus on issues that the end users have difficulty with. A heuristics evaluation (preferably with user involvement) is also important, but it inevitably considers the system as a collection of software entities.

A questionnaire such as SUMI represents the end result of a lot of research effort. The tester gets the result of this effort instantly when SUMI is used: the high validity and reliability rates reported for SUMI are to a large measure due to the rigorous and systematic approach adopted in constructing the questionnaire and to the emphasis on industry-based testing during development. However, as with all tools, it is possible to use SUMI both well and badly. Care taken in establishing the context of use, characterising the end user population and understanding the tasks for which the system will be used supports sensitive testing and yields valid and useful results in the end.

Heuristic evaluation and SUMI are testing techniques that can be applied to start usability testing or when limited resources for usability testing are available. Of course it is always a risk management decision, if usability is *thé* most critical succes factor, more thorough techniques such as full usability test should be applied. However, looking at current industrial usability practices, a large take-up of the discussed discount usability testing techniques would provide a great improvement for most projects and organisation ultimately leading to more usable and more user-friendly systems. User interfaces account for almost 50% of the code in modern software. In contrast, how much are you currently spending on usability testing?

Literature

Bevan, N. (1997), Quality and usability: a new framework, in: E. van Veenendaal and J. McMullan (eds.), *Achieving Software Product Quality*, Tutein Nolthenius, 's Hertogenbosch, The Netherlands

Bos, R. and E.P.W.M. van Veenendaal (1998), For quality of Multimedia systems: The MultiSpace approach (in Dutch), in: *Information Management*, May 1998

ISO/IEC FCD 9126-1 (2001), *Information technology - Software product quality - Part 1 : Quality model*, International Organization of Standardization

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl

ISO 9421-10 (1994), *Ergonomic Requirements for office work with visual display terminals (VDT's) - Part 10 : Dialogue principles*, International Organization of Standardization

ISO 9241-11 (1995), *Ergonomic Requirements for office work with visual display terminals (VDT's) - Part 11 : Guidance on usability*, International Organization of Standardization

Jacobson, I. (1992), *Object Oriented Software Engineering; A Use Case Driven Approach*, Addison Wesley, ISBN 0-201-54435-0

Kelly, M. (ed.) (1994), MUSiC Final Report Part 1 and 2: the MUSiC Projects, Brameur Ltd, Hampshire, UK

Kirakowski, J., The Software Usability Measurement Inventory: Background and Usage, in: *Usability Evaluation in Industry*, Taylor and Francis

Kirakowski, J. and M. Corbett (1993), SUMI: the Software Usability Measurement Inventory, in: *British Journal of Educational Technology*, Vol. 24 No. 3 1993

MultiSpace (1997), *Report on demand oriented survey*, MultiSpace project [ESPRIT 23066]

National Physical Labotory (NPL) (1995), *Usability Context Analysis: A Practical Guide*, version 4.0, NPL Usability Services, UK

Nielsen J. , (1993) *Usability Engineering*, Academic Press

Nielsen, J. and R.L. Mack (eds.) (1994), *Usability Inspection Methods*, John Wiley & Sons, Inc.

Preece, J. et al, *Human-Computer Interaction*, Addison-Wesley Publishing company

Rubin, J. (1994), *Hanbook of Usability Testing: How to Plan, Design, and Coduct Effective Tests*, John Wiley & Sons, Inc.

Trienekens, J.J.M. and E.P.W.M. van Veenendaal (1997), *Software Quality from a Business Perspective*, Kluwer Bedrijfsinformatie, Deventer, The Netherlands

The Author

Drs. Erik P.W.M. van Veenendaal CISA has been working as a practitioner and manager within the area of software quality for a great number of years. Within this area he specialises in testing and is the author of several books, e.g. "Testing according to TMap" and “Software Quality from a Business Perspective”. He is a regular speaker both at national and international testing conferences and a leading international trainer in the field of software testing. Erik van Veenendaal is the founder and managing director of Improve Quality Services BV (www.improveqs.nl). Improve Quality Services BV provides services in the area of quality management, usability, inspections and testing. He is a senior lecturer at the Eindhoven University of Technology and is also on the Dutch standards committee for software quality.

Improve Quality Services BV, Waalreseweg 17, 5554 HA Valkenswaard
Tel: 040-2089283, Fax: 040-2021450, E-mail: info@improveqs.nl