Evaluating the Usefulness and the Ease of Use of a Web-based Inspection Data Collection Tool

Oliver Laitenberger
Fraunhofer Institute for
Experimental Software Engineering
Sauerwiesen 6
D-67661 Kaiserslautern-Siegelbach
Germany
+49 (0)6301 707251
laiten@iese.fhg.de

Horst M. Dreyer
Mummert + Partner
Unternehmensberatung AG
Neue Weyerstrasse 6
D-50676 Köln
Germany
Horst.Martin.Dreyer@mummert.de

ABSTRACT

Software tools can only be effective if users accept them. Investigating user acceptance requires a model explaining people's attitudes and behaviour as well as reliable and valid measurement instruments. This paper applies Davis' technology acceptance model (TAM) and its questionnaire-based measurement instrument for evaluating a newly developed Web-based inspection data collection tool (WIPS) from a user's perspective. TAM postulates that two particular user beliefs, usefulness and ease of use, predict tool usage and, thus, tool acceptance quite well. To investigate this and to assess the usefulness and ease of use of WIPS, we performed an experiment with computer science students as subjects.

Our experimental results confirmed the reliability and the validity of the TAM and its questionnaire. Moreover, WIPS received high ratings for both TAM determinants. These results, together with the fact that usefulness was correlated to self-predicted future usage, explain why most of our subjects preferred WIPS over paper-based forms for inspection data collection.

The reliability and validity of the TAM questionnaire is not limited to inspection tools, but applies to the user acceptance evaluation of tools and techniques in general. We demonstrate this with data from an experiment we performed in an industrial setting with professional software developers to evaluate different inspection techniques.

1. Introduction

The collection of inspection data is considered essential for monitoring, controlling, and improving software inspections. During the various steps of the inspection process, the inspection participants are required to record defect and effort data on paper-based forms. In most cases, data are afterwards entered into a database for analysis. More recently, some authors have suggested that tool support be provided for automating software inspection [12], [15], [18]. Apart from other cited benefits a tool often facilitates data collection activities by replacing paper-based forms with tool-based forms. However, two issues must be resolved beforehand: First, if inspection participants regard a data collection tool as cumbersome, they will be reluctant to use it. Regarding inspection data collection, this attitude may result in invalid, unreliable, or incomplete inspection data. Second, several tools are currently available to support software inspection [1], [4], [12], [13], [14], [15], [18], and each tool has its merits. To evaluate and compare these tools, it is important to also consider the user's opinion apart from pure technical or functional criteria.

Both issues highlight the need to require a better understanding of people's attitudes and the "people" factor in general. The importance of the people factor has also been pointed out in a report of the National Research Council [17]: "... Recognizing and characterizing the human attributes within the context of the software process are key to understanding how to include them in system and statistical models." A better understanding of the people factor can be used to guide the development of the next version of the tool. However, when measuring people's attitudes about tools, one must rely on subjective measures for making judgements or draw conclusions since no objective measures tell a researcher that a tool is "good" or "bad" from the user's perspective. This requires a model or a theory explaining people's attitudes and behaviour as well as reliable and valid measurement instruments for it.

This paper presents such a model and a questionnaire-based measurement instrument for evaluating a newly developed Web-based inspection data collection tool. The model is the Technology Acceptance Model (TAM) introduced by Davis et al. [7]. The goal of the TAM is to provide an explanation of the determinants of technology acceptance. TAM posits that two particular user beliefs, usefulness and ease of use, are of primary relevance for technology acceptance behaviour, which is an important requirement for actual technology usage. For measuring usefulness and ease of use, Davis [6] developed a questionnaire and assessed its psychometric properties, that is, its reliability and validity, in a set of experiments. He found that, indeed, ease of use and usefulness have a big impact on the usage of management information systems as suggested by the TAM. Adams et al. performed a replication and corroborated the results of Davis' experiments [8].

We performed an experiment to investigate whether the TAM and its questionnaire also holds for characterizing and evaluating a newly developed Web-based Inspection Process Support tool (WIPS) for inspection data collection. Within the experiment, 24 students inspected a code module and used paper-based forms as well as WIPS for collecting defect and effort data. Once they had completed the code inspection, they answered 14 questions of the questionnaire that purport to measure usefulness and ease of use. Based on these data, we confirmed the reliability and validity of the TAM questionnaire. Since both usefulness and ease of use are correlated to self-predicted future usage, they can be considered determinants of tool acceptance behaviors. This explains why subjects in the experiment preferred WIPS over paper-based forms for inspection data collection.

The use of the TAM as well as of the questionnaire, however, is not limited to the evaluation of tools. It can rather be applied to software technologies in general. To illuminate this, we provide some empirical evidence from an ongoing industrial experiment into whether the questionnaire and its underlying model is reliable and valid for determining the usefulness and ease of use of software inspection techniques.

This paper is structured as follows: Section 2 presents the determinants of the TAM and a description of the questionnaire items. Section 3 introduces the WIPS tool. Section 4 presents the experiment we performed. Section 5 discusses the results of our experiment. Section 6 presents some limitations of the experiment and the results of an ongoing industrial experiment as extension of this research. Section 7 concludes.

2. Model of Usefulness, Ease of Use and Self-predicted Future Usage

Using tools offers the potential for improving many mature software engineering techniques, such as software inspection. Whether this is in fact true depends on many factors. An important one to consider is the "*people factor*". People, that is, users, tend to use or not use a tool according to the extent to which they believe it will help them perform their job better. We refer to this determinant as the perceived usefulness of a tool. However, even if users believe that a given tool is useful, they may, at the same time, believe that the tool is too difficult to use and that the performance benefits are outweighed by the effort of using it. This is best captured by the ease of use determinant. To define these determinants in more detail, we stick to the definitions given in [6]:

- *Perceived usefulness* is "the degree to which a person believes that using a particular system would enhance his or her job performance." This follows from the definition of the word useful: "capable of being used advantageously." Hence, a tool high in perceived usefulness is one for which a user believes in the existence of a positive use-performance relationship.
- *Perceived ease of use*, refers to "the degree to which a person believes that using a particular system would be free of effort." This follows from the definition of "ease": "freedom from difficulty or great effort". A tool that is easy to use is more likely to be accepted by users.

Usefulness and perceived ease of use are both important determinant of the Technology Acceptance Model (TAM) introduced by Davis et al. [7]. It is the goal of this model to provide an explanation of the determinants of technology acceptance. This is helpful not only for prediction but also for explanation, so that researchers and practitioners can identify why a particular technology may be unacceptable, and pursue appropriate corrective steps.

Since no objective measures are available for measuring usefulness and ease of use, we have to subscribe to subjective measures for which we apply a Likert or "summative" scale [16]. One may also use the term "linear composite" to designate such a scale. In brief, Likert scaling may be described in the following manner: A set of items, e.g. questions of a questionnaire, consisting of a set of statements is given to subjects. They are asked to respond to each statement in terms of their own degree of agreement or disagreement. In our case, they have to select one of seven responses: extremely likely, quite likely, slightly likely, neither, slightly unlikely, quite unlikely, extremely unlikely. A score is assigned to each response and the scores belonging to a particular determinant are combined so that subjects with the most favorable attitude will have the highest determinant score while subjects with the least favorable attitudes have the lowest determinant score. Table 1 and Table 2 presents the scale items that were considered for the usefulness and ease of use determinant. We present the complete questionnaire in the Appendix. The different items were originally proposed by Davis [6] who derived them theoretically from research on self-efficiency theory, research on the cost-benefit paradigm from behavioral decision theory, and research on the adoption of innovations. The items were selected based on several empirical studies and showed high reliability and validity properties. We used these items as a starting point and tailored them for the evaluation of WIPS, which we developed for inspection data collection.

Usefulness
U1 Using WIPS in my job would enable me to accomplish tasks more quickly (Quick).
U2 Using WIPS would improve my job performance (Job performance).
U3 Using WIPS in my job would increase my productivity (Increase productivity).
U4 Using WIPS would enhance my effectiveness on the job (Effectiveness).
U5 Using WIPS would make it easier to do my job (Makes job easier).
U6 I would find WIPS useful in my job (Useful).

Table 1: Scale items of the usefulness determinant

Ease of Use
E1 Learning to operate WIPS would be easy for me (Easy to learn)
E2 I would find it easy to get WIPS to do what I want it to do (Clear and understandable).
E3 My interaction with WIPS would be clear and understandable (Controllable).
E4 It was easy to become skillful using WIPS (Skillful).
E5 It is easy to remember how to perform tasks using WIPS (Remember).
E6 I would find WIPS easy to use (Easy to use).

Table 2: Scale items of the ease of use determinant

Since we performed the evaluation of WIPS with German students, we translated the questionnaire from the English to the German language. To avoid any language bias, a professional translator improved our translation to ensure the semantics of the items did not change. However, after running the experiment and debriefing the subjects, we realized that the subjects misinterpreted two items. These items were “Job Performance” in the usefulness scale and “Controllable” in the ease of use scale. Our subjects associated “Job Performance” with the task of detecting defects, and not with the tool support for inspection data collection activities. We attribute this to the translation of items. Regarding the item “Controllable” debriefing revealed that most of our subjects did not understand the meaning of this item. As a consequence, we decided to extract both items from analysis.

We are interested in how usefulness and ease of use impact user acceptance. According to Davis [6] “Perceived usefulness is a strong correlate of user acceptance and should not be ignored by those attempting to design or implement successful systems.” In our case, a user accepts WIPS if he or she is going to use it in the future (self-predicted future usage). This serves us as an indicator whether users prefer WIPS for inspection data collection rather than paper-based forms. For measuring self-predicted future usage we captured the two additional items presented in Table 3.

Self-Predicted Future Usage
UA1 Assuming WIPS would be available on my job, I predict that I will use it on a regular basis in the future.
UA2 I would prefer using WIPS to paper-based forms for performing inspections.

Table 3: Scale items of self-predicted future usage

3. A Web-based Inspection Process Support Tool (WIPS)

This section presents the Web-based Inspection Process Support Tool (WIPS) for inspection data collection. Before describing WIPS in more detail, we explain the software inspection process for which we provide tool support.

3.1 Software Inspection Process

In the past two decades, software inspection [9] has evolved into a mature practice in software engineering. Inspection participants usually follow a well-defined, four-step inspection process depicted in Figure 1.

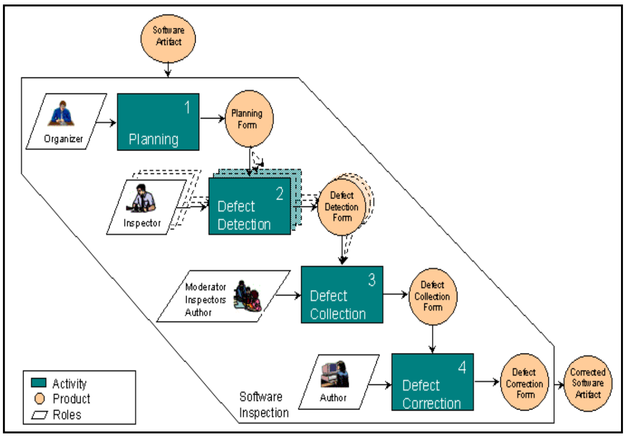

Figure 1: Overview of the inspection process

The author of a software artifact triggers the inspection process by submitting the software artifact for inspection. Throughout the planning step (1), an organizer is responsible for setting up the inspection, that is, assigning moderator and inspectors, making room arrangements, and distributing inspection material. During the defect detection step (2), inspectors prepare themselves individually, scrutinizing a software artifact for defects. As some of the defects suggested by an inspector may prove not to be real defects, inspection participants perform an inspection collection step (3). During this step the inspectors, a moderator and the author of the inspected software artifact meet with the purpose to collect the defects found by each inspector, to decide upon which of the potential defects are real ones, and to document the real defects. Besides, some additional defects may be detected and documented during an inspection meeting. Finally, the author of a software artifact is responsible for correcting all real defects collected during the collection step (4).

3.2 A Web-based Inspection Process Support Tool (WIPS)

Recently, some authors have suggested that the inspection process be automated at least partly [4], [15]. Data collection activities are candidates that can easily be supported by a tool. Hence, we implemented WIPS for inspection data collection. WIPS is built upon a client/server concept in which WIPS represents only one possible client application, that is, WIPS is a Web-based extension of a more general inspection tool IPS (Inspection Process Support). Since we have chosen the World Wide Web (WWW) as infrastructure, WIPS is independent of any particular system environment. WIPS provides Web-based point-and-click user interfaces based on Java applets. These applets are integrated in browser-based forms and allow inspection participants to access a database for storing inspection data (e. g., defect, effort) throughout the defect detection and defect collection step of an inspection. A major advantage of WIPS compared to paper-based forms is that the tool incorporates completeness checks, that is, the inspection participants are forced to decide upon the defect class and to give the location for each entered defect before a defect is stored in the database. For the defect detection and defect collection step of an inspection, the following two forms are provided:

(1) Defect Detection (Individual Preparation)
For each defect the location of the defect, the classification of the defect, and the description of the defect is collected. Moreover, an inspector must enter his/her defect detection effort (preparation effort) in minutes. Figure 2 depicts a screenshot of the detection form of WIPS.

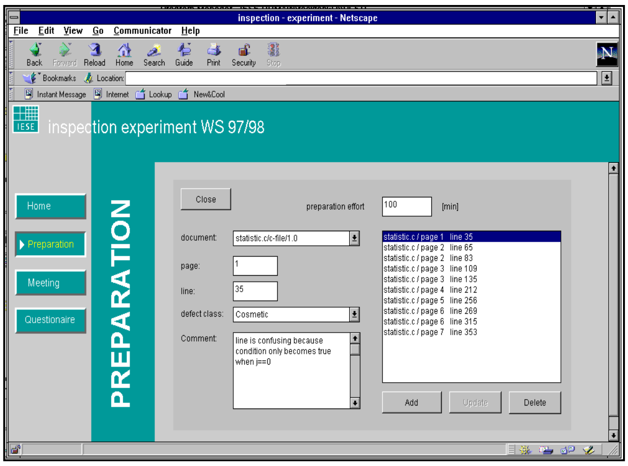

Figure 2: Screenshot of defect detection form

(2) Defect collection (Meeting)
Two or more inspection participants together perform an inspection meeting. Figure 3 shows the screenshot of the WIPS defect collection form. This form allows the inspection moderator to execute three different operations by simple point and click actions which avoids unnecessary typing: First, he or she can accept a potential defect as a real one. For this, all defects detected throughout the defect detection step are provided in the list box on the left-hand side. Once accepted as a real defect, the inspection moderator clicks on the potential defect highlighting it and then presses the "accept" button. Second, inspection participants can decide if different potential defects are referring to the same real defect. The inspection moderator can mark them as such, using the "=" button. Third, WIPS allows the inspection moderator to document defects detected during the inspection meeting (Add/Update button). For each defect the complete defect information is collected (location, defect class, comment). Moreover, the defect collection (meeting) effort is collected as well.

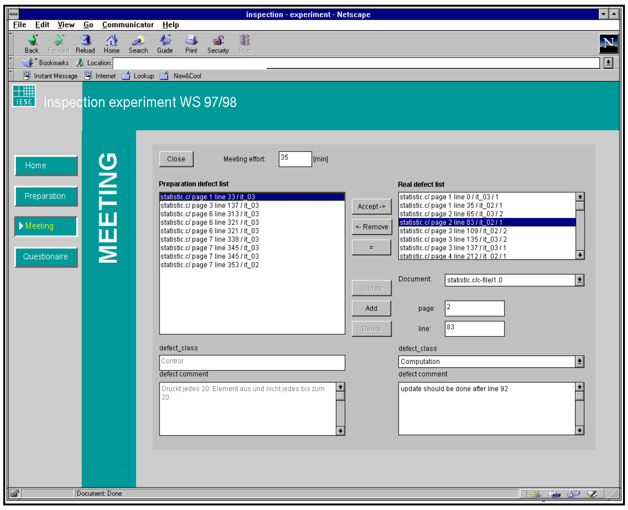

Figure 3: Screenshot of the defect collection form

4. Experiment

4.1 Goal of the Experiment

We performed an experiment because of two reasons. First, we wanted to validate the psychometric properties of the TAM and its questionnaire-based measurement instrument. Second, we wanted to know whether inspection participants prefer WIPS to paper-based forms for inspection data collection and whether the two TAM determinants usefulness and ease of use explain this preference. Only if the preference is demonstrated, WIPS can be considered an improvement of the software inspection process. Figure 4 gives an overview of the issues we are investigating.

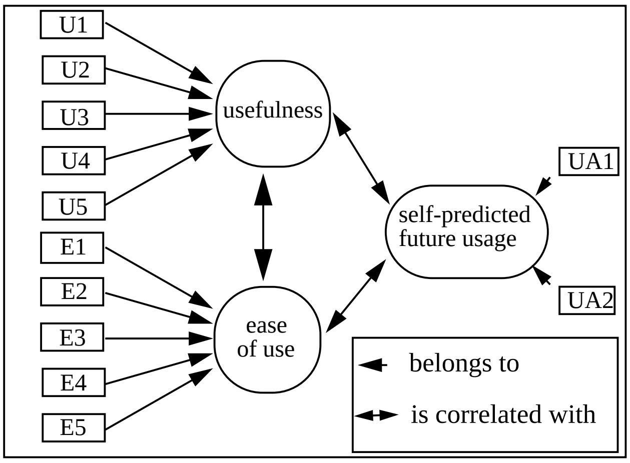

Figure 4: Model of usefulness, ease of use, and self-predicted future usage (TAM)

Regarding the psychometric properties, we do not have to ensure construct validity since the usefulness and ease of use questionnaire items are based on research on self-efficiency theory, research on the cost-benefit paradigm from behavioral decision theory, and research on the adoption of innovations. However, we have to check the reliability and factorial validity of the questionnaire items. Reliability concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials [4]. Factorial validity asks whether an item really belongs to a particular determinant or must be assigned to another. Assuming that our questionnaire is reliable and valid, we can determine the usefulness and ease of use of WIPS. This is a prerequisite to investigate whether these determinants really impact user acceptance (self-predicted future usage).

4.2 Variables

We are interested in three variables: perceived usefulness, perceived ease of use, and self-predicted future usage as defined previously. At this point, we need to emphasize that WIPS only supports inspection data collection. This tool does not support the individual defect detection activity of an inspection. Therefore, objective measures, such as the number of detected defects or the effort spent for defect detection, cannot be used for evaluating WIPS since inspection participants do not detect more defects or detect defects faster by using WIPS.

4.3 Subjects

The subjects of the experiment were graduate students of the Computer Science Department at the University of Kaiserslautern, Germany. They were enrolled in a software engineering course lasting a semester. This course teaches the basic software engineering principles. The course is supplemented by practical exercises. 24 students participated in the experiment. All students had their Vordiplom, an initial set of exams which students have to pass after at least two years of study and which includes theoretical, practical, and technical computer science, mathematics, and an elective class. Before participating in the experiment, the subjects filled out an experience questionnaire. The results of this questionnaire revealed that most subjects were familiar with using Web Browser technology.

4.4 Running the experiment

The experiment consisted of two steps: Throughout the first step, the subjects performed the defect detection and defect collection step of an inspection without tool support, that is, they collected all data on paper-based forms (2 hours). In the second step, we explained to the subjects how to use WIPS for inspection data collection (15 min.). The explanation reduced a possible learning effect, which may have biased the results. Then, the subjects performed the defect detection and defect collection step of an inspection using WIPS (2 hours). In our experiment one of the inspectors also performed the role of the moderator during the meeting. This is a viable inspection approach as shown by Bisant and Lyle [3]. Finally, the subjects filled out the usefulness/ease of use questionnaire.

5. Results and Discussion

5.1 Reliability

In most empirical studies, measures are rarely perfectly accurate (in the sense of stable or consistent). The degree of accuracy is called reliability. Roughly speaking, reliability is the extent to which one would obtain the same result if one were to administer the same measures again to the same person under the same conditions. The most widely used measure of reliability is Cronbach's alpha. Cronbach's alpha can be thought of as describing how much each measured item is correlated with every other item - the overall consistency of the test-, that is, the extent to which high responses go with highs and low responses go with lows across all items. Cronbach's alpha 0.8 is considered as highly reliable [5]. Table 4 shows the values of Cronbach's Alpha for each determinant:

	Cronbach's alpha
Usefulness	0.84
Ease of Use	0.82

Table 4: Cronbach's Alpha (Reliability)

Both the usefulness and the ease of use scales show high levels of reliability, since Cronbach's alpha is larger than 0.8 for both scales. Hence, the results demonstrate that the questionnaire is a reliable measurement instrument.

5.2 Factorial Validity

Factorial validity is concerned with whether the usefulness and ease of use items form distinct constructs. This can be checked with factor analysis. Factor analysis is often used when research has measured people on a large number of items. It tells a researcher which items tend to cluster together - which ones tend to be correlated with each other and not with other items. Each such cluster (group of items) is called a factor. We also use the term “determinant” instead of the term “factor”. Hence, factor analysis can be used to identify the underlying determinants of a set of items. The relative connection of each of the original variables to a factor is called that variable’s factor loading on that factor. Factor loadings can be thought of as the correlation of the item with the factor, and like correlations, they range from -1, a perfect negative association with the factor, through 0, no relation to the factor, to +1, a perfect positive correlation with the factor. While items have loadings on each factor, they will usually have high loadings on only one. A variable is usually meaningful to a factor if it has a loading of at least 0.7 [11]. But even lower values are sometimes considered important for a particular factor.

	usefulness	**ease of use**
Work More Quickly	0.77	0.37
Increase Productivity	0.83	0.07
Effectiveness	0.74	0.13
Makes Job Easier	0.73	0.10
Useful	0.75	0.38
Easy to Learn	0.09	0.84
Clear and Understandable	0.36	0.68
Easy to Use	0.56	0.63
Easy to Remember	0.20	0.69
Easy to become Skillful	0.10	0.85

Table 5: Factor analysis results

Table 5 shows the factor loadings. They indicate that the ten questionnaire items load on two different factors: usefulness and ease of use. However, some items in the ease of use scale have values slightly below 0.7. But since they load higher on the ease of use than on the usefulness determinant, we attribute them to the latter.

5.3 The Usefulness of WIPS

Figure 5 exhibits the usefulness results. The ratings of the summative results range between 26 and 32 (mean rating: 28.37). Considering that the maximum rating is 35, we can conclude that most of our subjects consider WIPS useful. To investigate this in more detail, Figure 5 also presents the results of each item in addition to the summative usefulness results. When asked directly about the usefulness of WIPS, our subjects considered WIPS to be very useful (mean of 6). Furthermore, they affirmed that WIPS makes their job clearly easier (mean of 6). Several reasons are responsible for this high rating: WIPS allows a subject to type in defect information and supports defect classification throughout the detection step of an inspection. In an inspection meeting, WIPS makes this information available, avoiding unnecessary typing.

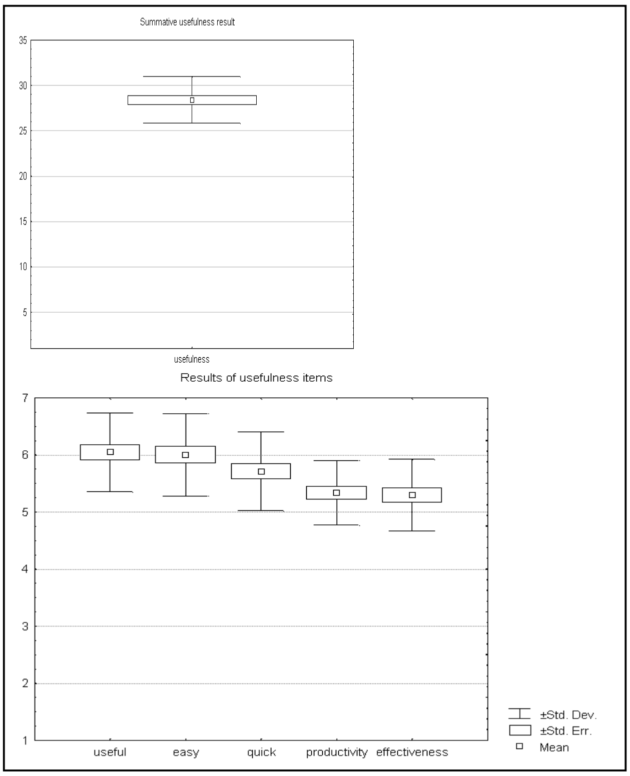

Figure 5: Usefulness results

Besides, users are not confronted with unreadable handwriting or unsorted defect lists. Hence, users confirm that WIPS makes their work more comfortable and is very useful for facilitating their work. The reduction of typing effort and the easy point- and click-facilities during the meeting are also reflected in the fact that our subjects affirmed that WIPS allows them to perform their job more quickly (mean of 5.5). The results of the productivity and effectiveness items reveal that most of the subjects are of the opinion that WIPS only slightly increases productivity and effectiveness. We found two possible explanations why the subjects scored these two items lower than the other ones: First, some WIPS users may have related their productivity and effectiveness judgment to the amount of code they can inspect within the given time slot. Since WIPS does not provide active support for a subject to scrutinize the code for defects, the use of WIPS does not impact the amount of inspected code. Hence, a subject scores lower on these two items. Second, although we applied a Web-architecture based on an application server that overcomes the performance limitation of the Common Gateway Interface (CGI), the response time of WIPS is sometimes rather low compared to a pure data base application because of the performance limitations of the WWW. In addition, it takes some time to start a Java applet since the applet code must be transferred from the server to the client system. Of course, waiting time may bias the attitude of a WIPS user such that he or she scores productivity and effectiveness lower than usual. Right now, we do not have results from evaluating the usefulness of other inspection tools. Hence, we consider our results a baseline for further inspection tool evaluation experiments or case studies.

5.4 The Ease of Use of WIPS

Figure 6 presents the ease of use results. The sum of the questionnaire items ranges between 27 and 35 (mean rating: 31.75). Regarding the maximum rating of 35, the ease of use rating can be considered high. Based on these findings, we can conclude that our subjects consider WIPS easy to use. A closer look at the different scale items reveals that learning to use WIPS is particularly easy for our subjects (mean of 6.6). Furthermore, participants consider WIPS very easy to use (mean of 6.3) and think that WIPS is clear structured (mean 6.29). We only found two items that had mean ratings slightly below 6: Skilful (mean 5.97) and remember (mean 5.8). But still, these values are positive judgment by our subjects.

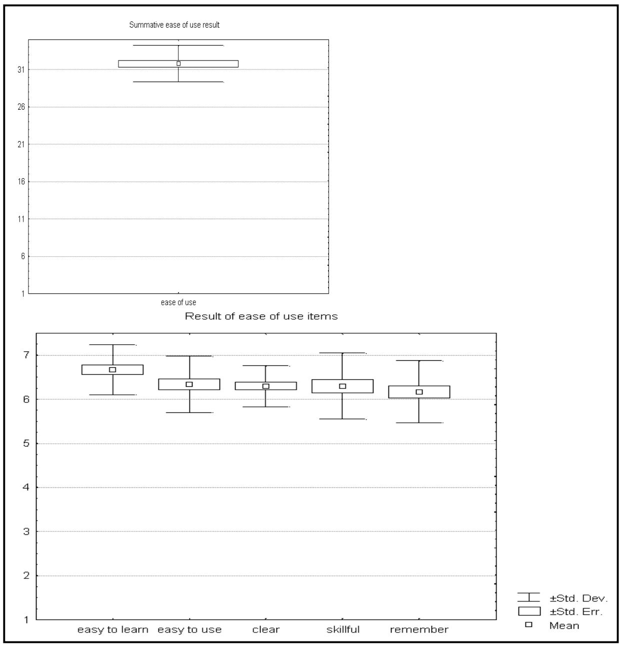

Figure 6: Ease of use results

We attribute the positive ease of use results to the design of the collection forms. Java-Applets allowed us to develop intuitive point-and-click user-interfaces. Hence, our subjects only needed little guidance on how to use WIPS for inspection data collection activities and thus considered WIPS easy to use

5.5 Correlating Usefulness, Ease of Use, and Self-predicted Future Usage

So far, we found that our subjects consider WIPS useful and easy to use. This does not tell us how their opinion impacts user acceptance. To investigate this in more detail, we correlated the summative results referring to the usefulness and ease of use scale to the summative results referring to self-predicted future usage. Since we assumed a correlation between usefulness and ease of use we did not apply regression analysis because of multicollinearity [2]. Table 6 shows the Pearson product correlation coefficients as well as the Spearman Rank Order Correlation coefficient in parentheses.

	Usefulness	**Ease of Use**	**Self-predicted future usage**
Usefulness	1	0.40 (0.32)	0.46 (0.41)
Ease of Use	0.40 (0.32)	1	0.27 (0.26)
Self-predicted future usage	0.46 (0.41)	0.27 (0.26)	1

Table 6: Correlation between usefulness, ease of use, and self-predicted future usage

Both ease of use and usefulness are positively correlated with self-predicted future usage. However, the correlation coefficient between usefulness and self-predicted future usage is much higher (0.46/0.41) than the correlation coefficient between ease of use and self-predicted future usage (0.27/0.26). One meaningful explanation for this finding is that users are driven to adopt a tool primarily because of the functions it performs for them, and secondarily for how easy or hard it is to get a tool to perform those functions. In a sense, usefulness mediates the effects of ease of use on self-predicted future usage. This explanation is also supported by the relatively high correlation coefficient (0.40/0.32) between ease of use and usefulness. To investigate whether the correlation coefficients are due to chance, we performed a correlation test as explained in [10]. With respect to the Pearson Product Moment Correlations coefficients, we found p-values of p=0.03 for the correlation between usefulness and self-predicted future usage and p=0.05 for the correlation of ease of use and self-predicted future usage. For the Spearman Rank Order Correlation

Coefficients, we found p-values of p=0.1 for the correlation between usefulness and self-predicted future usage and p=0.04 for the correlation of ease of use and self-predicted future usage. These p-values indicate that the correlations are not due to chance.

Although causality is not implied by the statistical analysis performed, these results show that usefulness and ease of use are important determinants that impact and explain self-predicted future usage. Most of our subjects consider WIPS useful and easy to use. Moreover, they are going to use it in the future, that is, subjects prefer tool-based inspection data collection over paper-based inspection data collection.

6. Limitations and Extension

As in any empirical study, we need to discuss the limitations of the experiment. First, our findings are tied to the chosen inspection process. In the context of our experiment, participants individually looked for defects and then performed a classical face-to-face meeting. Some authors suggest variations of this process such as replacing the classical face-to-face meeting with a distributed, asynchronous discussion step [8] [10]. Since this inspection approach is only feasible with adequate tool support, the usefulness rating might be different there. Second, the usefulness and ease of use measures are based on self-reported questionnaire items as opposed to objectively measured ones. However, our results show that the questionnaire items are reliable and valid. Furthermore, there are no objective measures to capture usefulness and ease of use. Hence, the only possibility is to investigate the mechanisms driving user behavior with the help of subjective measures. We strongly believe that the "people factor" reflected in user behavior is an important one to consider while developing any particular tool or suggesting any new software engineering technique. This opinion is shared by many experts as reported by the National Research Council [17]. However, this factor has often been neglected in software engineering research and practice. One reason might be the lack of valid and reliable measurement instruments. This research provides one step to overcome this obstacle. Finally, the subjects of the experiment were students. Experiments with students are usually characterized by high internal but low external validity. This limits our possibility to generalize our findings.

The model as well as the questionnaire is not limited to the evaluation of inspection tools. It is also applicable for the evaluation of software engineering techniques. As an extension of this research, we performed an experiment with professional software developers in an industrial setting. The goal of the experiment was an evaluation of different inspection techniques. If our measurement instrument is reliable and our model valid, we can understand the subjective factors that impact a developer's opinion. This understanding is very valuable to improve a given inspection technique, that is, to make it more useful and/or easier to use and, hence, more viable for industrial use.

As part of the experiment we used our proposed questionnaire to let participants subjectively evaluate the usefulness/ease of use of a particular inspection approach. Hence, we investigated the technology acceptance model. In this case we considered self-predicted future usage even more important than in the student's experiment because new approaches to improve software quality, such as a particular inspection technique, do not work unless software developers consider it useful and easy to use.
After analysing data from more than 60 professionals, the results in Table 7 show that for this experiment in an industrial setting with a different object of study (i.e., inspection techniques) the questionnaire is a reliable measurement instrument. Factor analysis of the experimental data confirms that the questionnaire items discriminate between the usefulness and the ease of use determinant.

	Cronbach's alpha
Usefulness	0.86
Ease of Use	0.88

Table 7: Cronbach's Alpha (Reliability)

Table 8 depicts the correlation coefficients between usefulness, ease of use, and self-predicted future usage of inspection techniques. They are even higher than the ones we found in the experiment with students. This finding corroborates the TAM and its measurement instrument.

	Usefulness	**Ease of Use**	**Self-predicted future usage**
Usefulness	1	0.51	0.65 (0.60)
Ease of Use	0.51 (0.52)	1	0.49 (0.48)
Self-predicted future usage	0.65 (0.60)	0.49 (0.48)	1

Table 8: Correlation between usefulness, ease of use, and self-predicted future usage

7. Conclusions and Lessons Learned

In this paper, we presented Davis' technology acceptance model and its questionnaire-based measurement instrument for evaluating the usefulness and ease of use of WIPS, a Web-based inspection tool supporting inspection data collection. The results of an experiment show that the TAM questionnaire is reliable and discriminates between the usefulness and the ease of use determinant. The subjects of our experiment considered WIPS useful and easy to use, which explains why the subjects prefer WIPS for inspection data collection to paper-based forms. At this point we have to add that we, as experimenters, consider WIPS very useful since we were not confronted with unreadable handwriting of subjects or missing data (thanks to the consistency checking of WIPS). These advantages, together with the fact that data are already stored electronically, significantly decreased our analysis effort.

We encourage other researchers to use the questionnaire as a starting point for evaluating the usefulness and ease of use of their (inspection) tools from the user's point of view. We also provided quantitative evidence that the questionnaire is also applicable for evaluating the usefulness and ease of use of software engineering techniques in general. To this end, we succinctly summarize our experiences and lessons learned in a way that may help improve the questionnaire for future empirical studies:

- **Adapt scale items carefully**
  The questionnaire items must be carefully adapted to the given context. One must be aware that changing expressions may result in different interpretations by subjects and, hence, in different results.
- **Be as precise as possible**
  Some expressions, such as productivity or effectiveness, are often ambiguous. Hence, they need to be defined more precisely.
- **Omit the middle alternative ("neither"- value)**
  Students or subjects in experiments in general are often not evaluated based on their opinion. Hence, they must not fear any consequences of their voting. We, therefore, recommend to remove the middle value from the questionnaire items, that is, reduce the item scale from 7 to 6 points, since the middle value ("neither"-value) often does not provide valuable information about the direction in which some subjects lean.
- **Mix items referring to different determinants**
  When tool users are confronted with similar items they tend to overestimate the meaning of single words and search for differences among them. In the original studies, the different scale items were hidden in a questionnaire consisting of more than 60 questions. Hence, we recommend mixing the scale items with other questions, e.g., questions about evaluating an experiment or, at least, mix them up.
- **Ensure data completeness and data consistency using an on-line questionnaire**
  We provided the questionnaire on-line. This helped us ensure completeness and consistency of the questionnaire data.

Regarding future work, we plan to extend WIPS by providing on-line document handling (colored keywords, comfortable navigation facilities). Furthermore, we intend to investigate the usefulness and the ease of use of WIPS in an industrial context. Finally, we plan to apply the TAM and its questionnaire to further investigate the acceptance of software engineering techniques in general.

Acknowledgement

We are grateful to Colin Atkinson, Sonnhild Namingha, and the anonymous reviewers for their valuable comments on this paper. In addition, we thank Dieter Rombach, Ralf Reussner, and Wolfgang Scholz as well as the students for their contribution to the experiment.

References

[1] F. Belli and R. Crisan. Towards automation of checklist-based code-reviews. In International Symposium on Software Reliability Engineering, 1996.

[2] W. D. Berry and S. Feldman. Multiple regression in practice. Sage Publication, 1985.

[3] D. B. Bisant and J. R. Lyle. A two-person inspection method to improve programming productivity. IEEE Transactions on Software Engineering, 15(10):1294–1304, October 1989.

[4] L. Brothers, V. Sembugamoorthy, and M. Muller. Icicle: Groupware for code inspection. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, pages 169–181, 1990.

[5] E. G. Carmines and R. A. Zeller. Reliability and validity assessment. Sage Publication, 1979.

[6] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, pages 319–340, 1989.

[7] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8):982–1003, august 1989.

[8] P. A. Todd Dennis, A. Adams, and R. R. Nelson. Perceived usefulness, ease of use, and usage of information technology: A replication. MIS Quarterly, 6(3):227–247, june 1992.

[9] M. E. Fagan. Design and code inspections to reduce errors in program development. IBM Systems Journal, 15(3):182–211, 1976.

[10] P. Cohen and J. Cohen. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Inc., Publishers, 1983.

[11] C. W. Mueller and J. Kim. Introduction to factor analysis, what it is, and how to do it. Sage Publication, 1978.

[12] P. M. Johnson and D. Tjahjono, An Instrumented Approach to Improving Software Quality through Formal Technical Review, in Proceedings of the Nineteenth International Conference on Software Engineering, pp. 118-127, IEEE Computer Society Press, May 1997.

[13] R.F. McKenney J.W. Gintell, and M.B. Houde. Lessons learned by building and using scrutiny, a collaborative software inspection system. In Proceedings of the Seventh International Workshop on Computer-Aided Software Engineering, pages pp. 350–357, 1995.

[14] J. C. Knight and E. A. Myers. An improved inspection technique. Communications of the ACM, 36(11):51–61, 1993.

[15] F. Macdonald, J. Miller, A. Brooks, M. Roper, and M. Wood. Automating the software inspection process. Technical report, Empirical Foundations of Computer Science (EFoCS), University of Strathclyde, UK, 1996.

[16] J. P. McIver and E. G. Carmines. Unidimensional scaling. Sage Publications, 1991.

[17] Panel on Statistical Methods in SoftwareEngineering. http://www.nap.edu/readingroom/books/statsoft/, 1993.

[18] D. E. Perry, A. Porter, L. G. Votta, and M. W. Wade, Evaluating Workflow and Process Automation in Wide-Area Software Development, in Proceedings of the Fifth European Workshop on Software Process Technology (Carlo Montangero, ed.), Lecture Notes in Computer Science Nr. 1149, (Berlin, Heidelberg), pp. 188-193, Springer-Verlag, Oct. 1996

Appendix

Perceived Usefulness

Q1: Using WIPS in my job would enable me to accomplish tasks more quickly.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q2: Using WIPS would improve my job performance.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q3: Using WIPS in my job would increase my productivity.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q4: Using WIPS would enhance my effectiveness on the job.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q5: Using WIPS would make it easier to do my job.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q6: I would find WIPS useful in my job.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Perceived Ease of Use

Q7: Learning to operate WIPS would be easy for me.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q8: I would find it easy to get WIPS to do what I want it to do.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q9: My interaction with WIPS would be clear and understandable.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q10: It was easy to become skillful using WIPS.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q11: It is easy to remember how to perform tasks using WIPS.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q12: I would find WIPS easy to use.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Self-predicted future usage

Q13: Assuming WIPS would be available on my job, I predict that I will use it on a regular basis in the future.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	

Q14. I would prefer using WIPS to paper-based forms for performing inspections.

likely								unlikely
	extremely	quite	slightly	neither	slightly	quite	extremely	