



HAL
open science

Information Artifact Evaluation with TEDSrate

Hans J. Scholl, William Menten-Weil, Tim S. Carlson

► **To cite this version:**

Hans J. Scholl, William Menten-Weil, Tim S. Carlson. Information Artifact Evaluation with TEDSrate. 16th International Conference on Electronic Government (EGOV), Sep 2017, St. Petersburg, Russia. pp.359-377, 10.1007/978-3-319-64677-0_30 . hal-01702986

HAL Id: hal-01702986

<https://inria.hal.science/hal-01702986v1>

Submitted on 7 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Information Artifact Evaluation with TEDSrate

Hans J. Scholl, William Menten-Weil, and Tim S. Carlson

University of Washington, Seattle, WA, United States

{jscholl, wtmenten, timca}@uw.edu

Abstract. The evaluation of systems or artifacts as “outcomes” of software engineering (SE) projects has been a focus of study in SE-related research for quite some time. In recent years, evaluating artifacts, for example, mobile applications or websites has become more important, since such artifacts play increasingly critical roles in generating revenues for businesses, and the degree of artifact effectiveness is seen as a competitive factor. With the TEDS framework/procedure a novel and comprehensive approach to systematic artifact evaluation and comparison had been presented a few years ago, whose effectiveness and analytical power in comprehensive and highly detailed artifact evaluations and comparisons was empirically shown; however, despite its demonstrated capability TEDS still proved to be time and resource consuming like other evaluation approaches before. In order to overcome these constraints and provide evaluative feedback more quickly to developers and service providers, TEDSrate, a Web-based evaluation tool employing the TEDS framework/procedure, was developed. The tool was tested with two real-world organizations, the City of Seattle Emergency Operations Center (EOC) and the Seattle Sounders Football Club. The tests suggest that the highly configurable TEDSrate tool can fully implement and administer the TEDS framework/procedure and, at the same time, provide instantaneous, cost-effective, comprehensive, and highly detailed artifact evaluations to both developers and service providers.

Keywords: TEDS Framework and Procedure; TEDSrate; Information Artifact Evaluation; Information Artifact Comparison; Usability Studies; Value Added Criteria, Government Websites, Government Apps.

Introduction

Assessing the aptitude and appropriateness of software systems relative to both purpose and requirements along with evaluating their performance relative to user expectations has been a recurring theme in software engineering research for a long time. Investigations in these areas intend to contribute to improving overall system design and support the initial development and further evolution of an artifact in use so that systems better match purpose, requirements, and users' expectations. In a more general sense, such studies help to better understand the factors, which lead to software engineering success. However, for reasons of high cost,

heavy time commitments on part of both developers and user-evaluators, and institutional barriers among other hindering factors, systematic software artifact assessments and evaluations have been found difficult to conduct persistently (Buse, et al., 2011).

Furthermore, numerous aspects have to be considered when assessing and evaluating software systems ranging from internal architecture and code efficiency investigations, over studies on the effectiveness of human-computer interaction to user satisfaction and usability among others so that the purposes and foci of evaluative studies can vary widely. What constitutes ultimate software engineering success, hence, is still an open debate (Ralph and Kelly, 2014). As shown in the next section, user satisfaction and effective use-related studies have been conducted in increasing numbers in recent years; however, criteria and frameworks used in such studies are also of a wide variety making it difficult to compare study results.

Interestingly, in times of burgeoning mobile and web-based applications (apps), which compete for market share, evaluative user satisfaction and effective-use studies have rarely been used to compare such artifacts, which could greatly help ongoing software engineering efforts in such markets. A few years ago, the TEDS framework and procedure was introduced (Scholl, et al., 2011) and successfully utilized in a number of empirical user satisfaction and effective-use studies, which also encompassed detailed artifact comparisons [10, 19, 20, 22].

While TEDS has demonstrated its effectiveness and analytical power in these studies leading to highly detailed and comprehensive results, it nevertheless also demonstrated its limitations with regard to the aforementioned constraints of high cost, heavy time commitment, and difficulties in user-rater/evaluator recruitment. In order to address and mitigate these three specific barriers, the researchers developed, introduced, and tested TEDSrate, a Web-based application (app), which allows recruiting and employing user-rater/evaluators anytime and everywhere. In this paper, TEDSrate, its uses, and the initial experiences with using it in evaluative studies, are presented and discussed.

The paper is organized as follows: In the next section, related work is reviewed leading towards the research question. Then, the design of TEDSrate is presented followed by the description of real-world pilot tests of the application. The results of the pilot tests are discussed, followed by the presentation of future work building on this discussion. The paper then concludes that frameworks/ procedures like TEDS and supporting applications such as TEDSrate can effectively help conduct systematic user satisfaction and effective-use studies.

Literature Review

As mentioned before, determining and measuring the ultimate success of a software engineering project and its resulting artifacts has been a focus of debate for a long time. Already in the early 1980s fairly detailed categories had been specified for determining and assessing the relative value added by information sys-

tems regarding the specific contexts of their use and the respective information environment, in which they operate (Taylor, 1982). Later, the DeLone & McLean (D&M) model of information system success in its various evolutionary versions (Delone and McLean, 2003; DeLone and McLean, 1992) has served as a reference on a high level of abstraction in a number of SE-related fields and subfields (Ralph and Kelly, 2014; Seddon, 1997). The D&M model basically relates three high-level variables of quality (information quality, system quality, and service quality) to equally high-level variables of system use (or, the intent of its use) and the user satisfaction, which in turn are said to lead to measurable or perceived net benefits, which feed back on system use and user satisfaction, the latter two of which are also connected via feedback (Delone and McLean, 2003). Addressing these feedback relationships another recent study pointed at the importance of project efficiency, artifact quality, market performance, impact on stakeholders, and time as influential dimensions of software engineering success (Kitchenham, et al., 2005; Ralph and Kelly, 2014). Software engineering success along with overall information system or artifact success apparently depends on interacting and interdependent variables (Hurtado, et al., 2015), which render the respective outcomes to factors not completely controllable by designers, developers, and project leaders.

As a result, multiple studies focused on better understanding these context-related factors and feedbacks. For example, recent workshops and studies emphasized user involvement in design and testing [3, 4, 14, 24]. Others highlighted the importance of continuous feedback on artifact (use) performance [1, 17]. Yet, others have relied on built-in monitoring and self-tuning functionalities as well as automatic user review scanning and salient-issue ranking methods [5, 6, 13]. Also, although not new, recent studies have reintroduced the utilization of personae and scenarios in both artifact design and artifact evaluation (Anvari, et al., 2015; Schneidewind, et al., 2012).

However, the D&M model variables can hardly be studied in isolation, nor can they be effectively addressed when just employed on a high level of abstraction when it comes to design-relevant and artifact-specific recommendations (or comparisons). The TEDS framework and procedure (Scholl, et al., 2011), which represents a substantial extension to the aforementioned “Value-added Processes” work advanced in the 1980s (Taylor, 1982), not only breaks down into detail the six high-level variables of the D&M model, but also accounts for the interaction between the variables within a given context by employing the concepts of personae and scenarios. The TEDS framework distinguishes six major categories of (a) ease of use/usability, (b) noise reduction, (c) quality, (d) adaptability, (e) performance, and (f) affection. These main categories are further broken down into 40 sub-categories further specifying and detailing the main categories. The TEDS procedure, then, specifies thirteen steps of evaluating what is called an “information artifact,” which, as a summary term, is used to represent any information technology or software artifact that a human actor may use for her or his purposes within a certain context. The term “information artifact” encompasses “both sources and pieces of information as well as information systems and other information technology

artifacts” (Scholl and Carlson, 2012’, p. 141). The concept acknowledges that “information” is a context-dependent entity providing a certain meaning in the eyes of a beholder, and technology carrying and containing this very information can no longer sharply be distinguished from each other.

As mentioned, the TEDS framework and procedure has demonstrated its analytical power in various empirical studies [19, 20, 22], in which it was able to help derive detailed recommendations for developers and designers, and it also provided valuable competitive information to service providers who intended to improve their online offerings. However, while the results quite strongly proved the effectiveness and the overall concept of information artifact evaluation by means of the TEDS framework and procedure, it was still subject matter experts who had to carry out the detailed assessments and evaluations in a rather time-consuming and costly fashion (Buse, et al., 2011) and also in geographically limited areas, all of which would present serious constraints for the future use of TEDS.

Research Question and Methodology

As a natural next step, the authors considered building a web-based tool for using the TEDS framework and procedure, which would reliably facilitate the issuance of artifact assessments and evaluations to both subject matter experts and laypersons alike on a broad and potentially global scale. With increasing sample sizes and controllably established demographics, it was reasoned that this would enable information artifact evaluations rather inexpensively while comprehensively at the same time. In the following, requirements, design criteria, and design options for a web-based tool enabling the use of the TEDS framework and procedure are discussed. This addresses the research question:

RQ: *What kind of Web-based tool can help subject matter experts and laymen alike perform TEDS-based evaluations capably and with global access?*

Design Considerations

Overall Requirements. When analyzing how TEDS was “manually” used in projects of empirical information artifact studies, that is, when the projects followed the 13-step procedure as described elsewhere (Scholl, et al., 2011) without the use of information system technology (ICT) support, the authors identified functional and non-functional requirements of a to-be ICT-supported TEDS tool.

Functional Requirements

Rating/Evaluation Component: The TEDS tool had to be able to input, record, and display scale ratings (for example, on a 1-5 Likert scale) from human raters for up to six main categories and up to forty sub-categories of TEDS in a pre-specified number of scenarios and for a pre-specified number of personae. As part of the evaluation component the TEDS tool had to further be able to calculate and present/print average scale ratings per category/sub-category for each persona and scenario along with the standard deviation. Beyond recording numerical scale val-

ues the TEDS tool had to be able to record free-format text comments along with screenshots of a rated artifact for each category and sub-category in any persona-scenario couplet. Recording the ratings needed to occur in an IRB acceptable and human subjects protecting space along with online raters' detailed demographic information. The TEDS tool report component had also to be able to pivot results along each dimension. It also had to be able to include raters' comments and screenshots in reports. Rater-provided screenshots and comments had to be searchable/findable per artifact, scenario, persona, and rater.

Administration/Configuration Component: In order to make the TEDS tool usable for multiple projects and studies, a configuration tool was required; also, for the analysis of results an administration tool for projects and configuration was needed. The TEDS tool admin/configuration had to be able to freely configure categories and sub-categories (all, sub-sets, or extensions). It also had to be able to cluster and re-cluster sub-categories. The TEDS tool admin/configuration further had to be able to add, modify, and remove artifacts, scenarios, and personae. It had to be able to modify the descriptions of categories, sub-categories, and topical clusters. It had to be able to add, modify, and delete collected rating data. For use with external tools rating data and reports had to be exportable into CSV format. The export or handover to other utilities such as the R project for statistical computing had to be provided for post-processing of results.

Non-functional Requirements

For reaching out to expert and layman raters without geographical and time constraints, the TEDS tool needed to be Web-based and work on any Web browser. The browser-based user interface had to be easy to navigate and operate. For easy and straightforward rating and recording, the TEDS tool had to be able to display the information artifact to be rated without interfering with the artifact's functionality alongside the rating tool in a browser window. Given the electronic mass recruitment of raters, for example, via Facebook advertisement, the rater population would be diverse, and so would be their devices and platforms. Consequently, TEDS tool had to be able to support a wide range of devices. The user interface of the TEDS tool had to be adaptable and adjustable depending on the artifact under evaluation, for example, for mobile applications versus web pages, or for full-blown TEDS evaluations versus subset evaluations. Demographic questions had to be configurable relative to the respective TEDS study design. Ratings were to be recorded instantaneously. Rating sessions were to be able to be temporarily suspended and resumed at a later point in time without the loss of data. Raters were to be informed about the progress of the rating exercise relative to completion. Rating results were to be searchable instantaneously. High standard deviations in ratings along with other outliers were to be made visible. Graphics and charts were to support the analysis of rating results. Finally, recruiting and signing up raters, conduct-

ing ratings, recording and storing large amounts of data were to be performed in a fashion allowing for comprehensive empirical studies with low or no budgets

Design Criteria

When reviewing and considering the requirements, it quickly became clear that publicly available and generic tools such as Google Forms or SurveyMonkey were no suitable solutions for meeting IRB requirements and human subject protection needs and/or would carry prohibitively high price tags when signing up raters. Also, for the inaccessibility of respective data, statistical analyses on raters' demographics would have required significant overhead when using those generic tools. Furthermore, some essential functionality along with the need for flexible and robust configurability options would not have been attainable with such publicly available tools. Consequently, the researchers decided to build a homegrown tool, which would meet all requirements including the storage of collected data on secure institutional servers. Moreover, it was reasoned that a homegrown tool would far better fit the flexibility and configurability needs of future TEDS-based empirical projects.

Design Options

When analyzing various (also alternative) tool design options, we ultimately settled on utilizing the LAMP (Linux, Apache, MySQL, and PHP) stack. In our reasoning, while LAMP was popular, cost effective, and open source, it also provided the advantages of known runtime robustness along with generally high performance, global resource and support bases, excellent documentation, and sustainability for future development. Along these lines the high potential for continued future talent recruitment from a vast pool of knowledgeable developers for this platform was another important argument in favor of LAMP.

Among other options considered were Windows as server platform, noSQL as database, .Net as alternative to PHP, and native code development as opposed to Web-based application (app) development. In each single area as well as for the whole platform, we concluded that LAMP was favorable. Windows as proprietary server platform appeared more costly in terms of available development resources, installation cost, and upgradeability/version sustainability. The enterprise-grade .NET framework seemed to be overkill relative to the foreseeable present and future research needs of the envisioned relatively small system, which were seen as fully covered via PHP, the latter of which also provided rapid prototyping and app development along with boilerplate constructions of Web-based application program interfaces (APIs). Also, we did not expect much server-side logic to be needed. As a result, we saw PHP as a right-size/right-weight choice. On the client side, we could have opted for developing a native application instead of using a Web-based application. However, this would have led to a proprietary and high load of custom development and maintenance along with portability issues among others, whereas a Web-based client would be easier to develop, maintain,

and distribute. Finally, relational characteristics are a mainstay of TEDS-based use and usability studies so that a relational database concept was the natural choice over non-relational concepts. Among relational databases, MySQL had advantages of cost effectiveness, slimness, platform independence, robustness, and non-proprietaryness over other options such as Microsoft SQL Server, Oracle, or others. In summary, the LAMP stack appeared as a logical platform for the development and implementation of the Web-based TEDS rating tool, which was dubbed *TEDSrate*.

The TEDSrate Approach

According to the functional requirements, TEDSrate would need three main architectural components: (1) an administration and configuration component, (2) a rating or evaluation/assessment component, (3) a database component for storing study configurations as well as evaluation results and ratings along with qualitative data such as comments and screenshots, and (4) a result query and presentation component (see Figure 1). A fifth architectural component, that is, an automatic statistical post-processor was and still is under consideration for a future version of TEDSrate. In its current implementation, TEDSrate uses both plain php scripts and the object-oriented CodeIgniter (CI) PHP framework.

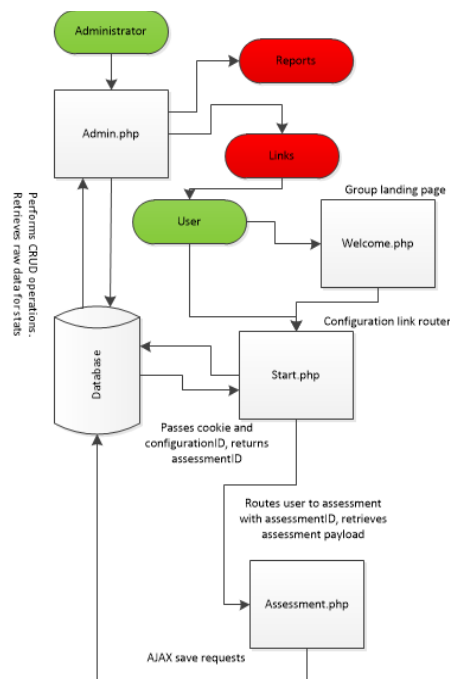


Figure 1 TEDSrate Overview

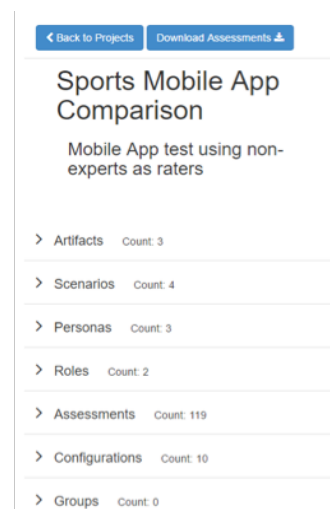


Figure 2 TEDSrate Admin/Configuration Tool (Project: Sports Mobile App Comparison)

The Admin/Configuration Component allows to create and manage TEDS research projects. On the server side, a new project is started in the admin function

by defining and attaching the project's use facets such as artifacts, personae, scenarios, and roles. Several scripts handle project setup and management including `adminproc.php` (for admin login/ logout), `start.php` (for handling the routing logic for new assessments and new users), `assessment.php` (a misnomer for legacy reasons, now containing an Angular template for issuing assessments), `upload.php` (for uploading rater screenshots and providing feedback to the raters), and `welcome.php` (for helping raters navigate configurations). In recent rewrites and updates to TEDSRate, CodeIgniter has been used as an efficient replacement method for previously used plain php models to interact with the data layer, since it also allows for the creation of a REST (representational state transfer) API, which is now the primary means of interacting with the database facilitating CRUD (create, read, update, and delete) operations on all entities of the data schema.

Furthermore, the Internal API handles specific processes such as receiving project overviews and generating report tables. On the client side, `Admin.js` is an Angular script, which supports the creation of project entities such as artifacts, scenarios, personae, roles, user interface configurations, and evaluations. `Admin.js` allows administrators to view rating results in the form of pivot tables presenting means and standard deviations. It further provides access to and graphically presents raters' demographic information. Moreover, `Admin.js` presents statistics along three dimensions: artifacts across a scenario, scenarios for one artifact, and an evaluation across a user interface configuration. The former two statistics provide aggregate data for the respective variables, the latter allows the granular inspection of individual evaluations when checking for data consistency and quality.

The Evaluation/Assessment Component. Much of the evaluation and assessment component resides on the client side, which has also mostly moved from legacy plain Javascript components to the Angular application module `Assessment.t.js`, which represents the logic for rater evaluations. This module is used for evaluations by both expert raters and layman raters and contains functionalities such as auto-saving, progress tracking, re-routing in case of evaluation/evaluator-rater mismatch, and screenshot uploading with progress feedback. The module also accounts for the various user interface configurations on the client side.

The Database Component. The relational database (see Figure 10 in the Appendix) contains tables for projects, artifacts, scenarios, personae, roles, and configurations. The latter serves as a container for four configuration types: attributes, assessment, questions, and user interfaces (UIs). It also provides an obscured ID in form of a hash, which allows raters to be added via the `start.php` script. Via attribute configuration, TEDS evaluation subsets can be configured (for example, instead of all forty sub-categories, only groups or clusters of categories/sub-categories can be selected for evaluation). The assessment configuration table specifies the key variables of the study, which are artifacts (usually a website or mobile app) and the scenarios, personas, and roles. The question configuration table serves as a target to associate the project with a group of survey questions. The UI configuration table contains the specification of the rating style (for example, Likert scale). The assessment table is the reference point for ratings, comments, and screenshots.

It also holds time stamp information. The attribute table specifies the TEDS category/sub-category or configured cluster. It further holds the attribute description or explanation in academic or layman language. The rating table stores the rating value for a single attribute. It also serves as the reference to attach attribute-related textual rater comments and screenshots. The question table holds the information on demographics questions (question title/name, description, and requirement status), whereas the response table stores the respective rater responses. Finally, the user table stores personal identifiers such as email address, first name, last name, and password along with the respective users' authorization level.

The schema also contains a number of associative entities such as project (parent), artifact, scenario, persona, role (children) or question (parent), project, artifact, scenario, persona, role, attribute (children).

Stored Procedures and Worked Scenarios. TEDSrate also contains about thirty stored procedures such as addPersona, addPersonaScenario, addProject, addProjectArtifact, addRating, addResponse, addScenario, addScreenshot, addUser, getAllArtifacts, getAllPersonae, getAllProjects, getCategories, getCriteria, getProject, getUser, updateCategory, and updateUser, among others.

Further, worked scenarios include starting a project, creating a configuration, and running a report.

Pilot Tests With Real-World Organizations

Concurrently, two TEDSrate-based evaluations of different artifacts were carried out, one of which in the environment of professional disaster response management at the City of Seattle's Emergency Operations Center (EOC), and the other with a major league soccer club (Seattle Sounders FC). In the case of the Seattle EOC a Web-based artifact was evaluated, which responders mainly work with on desktop computers during the response to an emergency or a disaster. In the other case, a mobile application was rated, which ticket holders, fans, and supporters of the Sounders FC franchise use to keep up to date about their team and to shop for franchise-related merchandise or tickets.

Government-internal Website Evaluation (WebEOC)

Intermedix' WebEOC® is a Web-based application suite, which is tailored to help Emergency Operations Centers (EOCs) manage the response to and early recovery from disasters. The suite is configurable and expandable and enjoys a relatively large user base among EOCs in the United States. In recent years WebEOC has been criticized for its cumbersomeness, complexity, and old-fashioned user interface.

The City of Seattle's EOC had a vested interest in identifying the exact problem areas of WebEOC from a user's perspective, that is, from a disaster responder's view. TEDSrate was configured and used to receive ratings and feedback from responders who had recently used WebEOC during a disaster response or exercise.

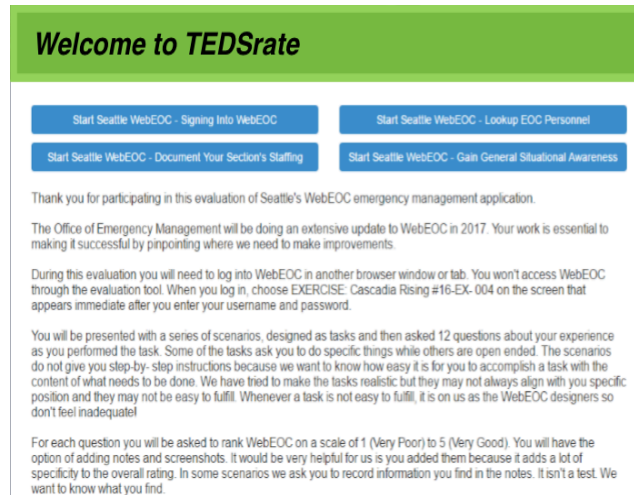


Figure 3 TEDSrate Configurable Entry Screen

In particular, four scenarios of utilization, each of which comprises one or more use cases, were seen as potentially in need of improvement along several lines (UI, performance, logic, etc.). The four utilization scenarios were (1) Signing into WebEOC, (2) Lookup EOC Personnel on duty, (3) Document Your Section's Staffing, and (4) Gain General Situational Awareness. The evaluation was carried out before and immediately after a major exercise was conducted involving over 200 responders in June 2016. The purpose of the evaluation was explained to responders on the entry screen (see Figure 3).

It is noteworthy that except for the introductory information in the entry page no further training of tool or method was required for responders to perform the requested evaluations for the four scenarios. The evaluation would be taken on a split screen, that is, the WebEOC artifact alongside the TEDSrate window.

APP Evaluation (SoundersS FC's Mobile iOS APP)

Almost every franchise in the US Major Soccer League (MLS) has implemented a mobile application for smart phones or notepads. While the websites of all franchises are designed, operated, and maintained by the League, the franchises have greater leeway to develop and implement their own mobile apps. The various MLS team websites are distinct in appearance (logos, team colors, etc.) and content (team-related information); however, they are uniform in terms of functionality and style guidelines. When it comes to mobile apps, the League appears to mandate only the adherence to guidelines of presentation style and merchandising, whereas the functionality of apps may widely differ between franchises.

Since its introduction to the League in 2009, Seattle Sounders FC has developed into a commercially highly successful MLS franchise with the far highest average attendance in the League (44,247 in 2015), which is more than double the

League's average (21,574 in 2015), and even exceeds the average attendance of the league with the highest attendance worldwide, that is, the German Bundesliga (43,177 in 2015) [12, 16].

A comprehensive TEDSrate-based evaluation of an early version of the second generation of the Sounders FC's mobile iOS app was conducted at a time, when the app development process had not concluded and was still open to extensions and modifications based on the evaluation results. The evaluation was performed in two rounds, first with expert raters who had been involved in a larger study, which had compared the mobile apps of a total of eleven leading professional soccer teams worldwide. The results of this separate study have been published elsewhere. These expert raters also evaluated the early second-generation mobile app of Sounders FC following the 13-step TEDS procedure in the traditional fashion without the support of TEDSrate. By mid-2015, the Sounders FC franchise agreed to collaborate with the research team upon organizing a TEDSrate-based evaluation of the second-generation mobile iOS app with the aim of incorporating the results of both experts' ratings and TEDSrate-based ratings in the further development of the app. Via targeted advertisements on Facebook "layman" raters were recruited who would then be directed to the TEDSrate evaluation site and asked to rate the second-generation Sounders FC mobile iOS app. As in the case of Web-EOC evaluation the "layman" raters would not receive any particular introduction nor training other than interactively available from the TEDSrate website. As intended the Facebook recruitments of "laymen" raters provided a wide spread of geographical, age, gender, and other backgrounds in the sample..

Demographics Module

When moving from purposively selected expert raters to a wider population of non-expert ("layman") raters it was imperative to collect demographic data in order to better quantify and qualify the results. More detailed and more specific demographic data would be needed for larger populations (for example, "Asian soccer fans," "North American soccer fans," or "European soccer fans", see Figure 5) than for smaller and more homogeneous populations such as "City of Seattle Emergency Responders" when making sense of and relating the rating results to demographic characteristics in the analysis phase.

As mentioned before, demographic questions are configurable accounting for larger and diverse populations.

The Rating Procedure

TEDSrate allows for configuring and adjusting the categories and sub-categories of the TEDS framework. As mentioned before, the framework consists of six main categories and forty sub-categories, which can be expanded or consolidated depending on the desired granularity of the specific evaluation project. In the case of "layman" evaluations fewer and consolidated categories/sub-categories serve the evaluation purpose more effectively than too specific and too detailed rating schemes, which typically only experts fully understand and then rate in an

informed fashion. We are referring to “experts,” in the context of TEDS, as individuals who have attended a TEDS framework and procedure training and, after completing an artifact rating, have also attended an inter-rater validity and consistency checking session.

Sports Mobile App Comparison - Player Information

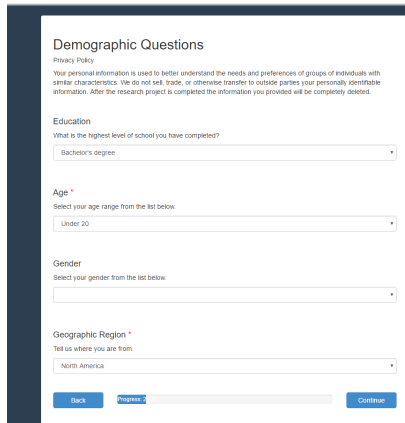


Figure 4 Sample Demographic Questions (Configurable)

Sports Mobile App Comparison Old - Mobile Access

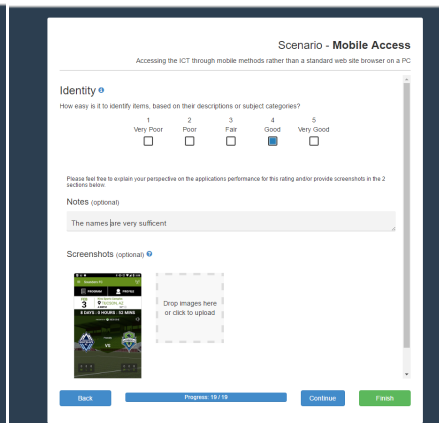


Figure 5 Sample Rating Screen with Likert Scale, Free-format Text Comments, and Screenshots (Configurable)

In the case of the WebEOC website evaluation as well as in the case of the “layman” evaluation of the second-generation Sounders FC mobile app a consolidated framework was used, which was reduced to twelve sub-categories (two for each main category—see sample screen in Figure 5), whereas the expert evaluation of the mobile app used the entire framework of forty sub-categories.

Transparent to the individual rater who uses the rating tool TEDSrate saves all data entries immediately via AJAX calls to the server. Each entry, whether it is a Likert scale radio button tick, a text comment, or an artifact screenshot is saved individually, so that client-to-server communications are relatively small and therefore fast.

Whatever configuration is used, the rater sees her advancement towards completion of the evaluation by means of a progress bar displayed at the bottom of the rating screen.

If raters have to postpone the completion of the evaluation for some reason, they find the latest data they had entered before pre-filled in the form, so that they can continue the rating exactly at the point, where they left it off.

Most artifacts are designed to serve multiple purposes and subsequently are used in practice in more than one scenario of utilization. However, the evaluation with TEDSrate has to distinguish between scenarios, since an artifact might be highly rated for some uses and certain scenarios, while it may fall short in others.

As an example, for the mobile apps of soccer clubs such as Sounders FC, Real Madrid, of FC Barcelona, the scenarios of “player information” and “schedule and results” might be evaluated among others. A rater, hence, has to go through the rating procedure as many times as separate scenarios were configured for evaluation. Once one scenario evaluation is completed, the rater needs to be reminded that other scenarios still need ratings. Once raters complete or leave a rating session unfinished, upon exiting the rating of a scenario, they are reminded of the overall completion status of their assignments (see Figure 6).

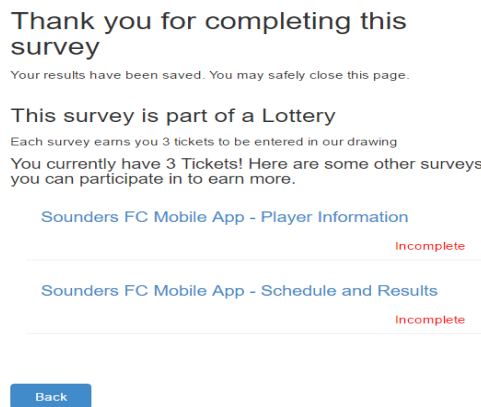


Figure 6 Survey Completion Update

In evaluation assignments with several pre-configured scenarios or attributes, TEDSrate also allows for the randomization of the order, in which the various scenarios or attributes are presented to the rater.

Mitigating Rater Fatigue

In the course of both artifact evaluations, the WebEOC website and the Sounders FC mobile app, rater fatigue was discovered. Some “layman” raters would leave the rating sessions behind incomplete even after repeated reminders. While the randomization of the order of assignments appeared to have already had some mitigating influence on rater fatigue, other means such as incentives were considered and became part of the TEDSrate tool during the practice test phase. In particular, when populations with potentially short attention spans are targeted, the incentive module can be configured. The implementation was performed in the format of a lottery, in which raters who completed the assignments would earn them “tickets” with certain material value, which could then be used for purchases or other benefits. In the case of the Sounders FC’s mobile app, the lottery-based mitigation strategy worked satisfactorily leading to much increased completion rates. The researchers also successfully experimented with giving out \$5 gift certificates to the first 25 raters who completed the TEDS surveys for two scenarios by using timestamp and user ID information. Likewise this led to more and faster completion of surveys in this particular pilot.

Presentation and Analysis of Results

In both pilot tests, the feature of the TEDSrate Admin utility, which lets the researchers track evaluations and lets them see even preliminary results in real time while the evaluations are still underway, was found highly informative and beneficial. All analytical functions can be performed this way, for example, inspecting pivot tables of ratings along the lines of configurations, scenarios, or artifacts, or after the evaluation project has ended. The utility also allows for selection and instantaneous analyses of demographic sub-samples, comment presentations, screenshot inspection, and data export to external analysis tools (for an example, see Figure 7).

	Player Information			Schedule and Results		
	Average	Std. Dev.	Count	Average	Std. Dev.	Count
Navigation and Findability	3.43	0.97	65	3.67	0.92	33
Structure	3.55	1.02	60	3.45	0.96	31
Identity	3.96	0.91	57	3.68	0.94	31
Substantiality	3.67	1.09	57	3.68	0.91	31
Completeness	3.91	0.88	56	3.71	1.04	31
Trustworthiness	4.04	0.99	56	3.90	0.98	31
Interaction	3.33	1.03	54	3.17	1.21	30
Customization	3.49	0.97	53	3.39	1.05	31
Savings	3.62	1.16	53	3.48	1.29	31
Confidence	3.79	1.13	52	3.81	1.14	31
Attractiveness	3.60	1.16	52	3.48	1.18	31
Enjoyment	3.44	1.07	52	3.42	0.99	31

Figure 7 Likert Ratings For Two Scenarios Along Twelve Sub-categories For the Sounders FC Mobile App

The visualization and formatting of results was found essential for analytical interpretation, also due to the sheer amount of detailed data, which was produced. Not only numerical data were target of visualization and formatted display but also comments, screenshots, and demographic information helping focus the analytical treatments and speed up the overall analysis process (for example, see Figure 8). In ongoing rating campaigns the immediacy of information availability, in particular,

with regard to demographic information helped target the rater recruiting so that the various identified personae could exactly be represented and matched by the sample of raters. Formatted displays for comments and screenshots supported the straightforward inspection of data and their analytical interpretation. When numerical data showed both relative strengths and weaknesses in a particular area, for example, “navigation and findability” in the scenario of “player information,” then the comments and screenshots, which raters had provided, could be inspected in that particular area (see Figure 7 and Figure 9).

Sounders FC Mobile App

Artifact Type: Website

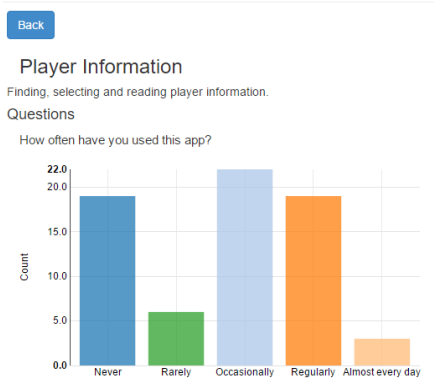


Figure 8 Visualization of Usage Frequencies in Support of Interpreting the Weight and Validity of Ratings

Ratings

<p>ID: 4930 Email: Value: 5.0 Comment: Very clean, easy to understand and responsive. The presentation of images as you drilled into player information is a real treat.</p> <p>Screenshots</p>
<p>ID: 4954 Email: Value: 3.0 Comment: Some navigation requires intuition or exploring, such as seeing stats from a past game (must click on game score). Also "Games" is a misleading term on the main menu.</p>
<p>ID: 4978 Email: Value: 4.0</p>

Figure 9 Inspecting Raters’ Comments and Screenshots in a Target Area Based on Clean and Formatted Displays

Discussion

As shown in the section on related work above, software engineering success depends on a number of interacting and interdependent variables, some of which escape the developers’ span of control, whereas others, which can be directly influenced, have so far gone unattended for the most part due to prohibitive cost and overwhelming commitment of resources and time needed to uncover deficiencies in, for example, artifact quality, attractiveness, user satisfaction, and system use among others.

Feedback, if any, which could practically and effectively influence how developers and designers tweak or reshape an artifact to better meet expectations and needs, would be slow in coming and probably incomplete. While the TEDS framework and procedure might be the most comprehensive and systematic analytical lens available for assessing, evaluating, and comparing artifacts, it also suf-

ferred from the high cost incurred, long time to conclude, and heavy resource commitment necessary in order to arrive at detailed, conclusive, and robust results. In many instances, however, even if such a level of effort had been expended, it would not have produced the needed feedback in due time, and, for example, market opportunity might have already vanished, or worse, damage had already been inflicted. The critical question then became how the prohibitive high cost, long turnarounds, and excessive resource commitments for systematic artifact evaluations could be cut down without compromising the validity and robustness of results. This led the research group to consider, specify, design, develop, and test TEDSrate in practice.

The tool underwent two real-world tests, one with the City of Seattle Emergency Operations Center (EOC) for a desktop-operated web-based application suite (WebEOC), which serves as the Center's linchpin in disaster response. The other real-world test was simultaneously conducted with Seattle Sounders FC for the soccer franchise's mobile application, which is the centerpiece of interaction between the club and its supporters and match attendees.

These two tests greatly demonstrated the effectiveness and utility of the tool, which produced robust and reliable results, which were used by both organizations to make targeted changes to the configuration of their respective artifacts. In the case of Sounders FC, the test identified in fine detail such areas that needed improvement. Moreover, informed by rater comments and screenshots and through pinpointed comparisons with other "best-in-class" implementations, detailed design recommendations were given to the mobile app developers, many of which have meanwhile been developed and implemented into version 2 of the Seattle Sounders FC mobile app.

The two tests were conducted over a period of six weeks. A total of 90 raters were involved, most of whom completed all Web-based TEDSrate surveys in all scenarios, to which they were assigned. The recruiting of "layman" raters was found easier, when certain material incentives were offered, for example, gift cards. Recruiting raters for the Seattle Sounders FC app via the Sounders' Facebook site by means of targeted Facebook advertisement was straightforward. In the case of WebEOC, the raters were recruited via EOC-internal email invitation. However, in other artifact evaluation and comparison studies, different recruiting approaches may also be effective.

Since TEDSrate works web-based, the reach of this artifact evaluation and comparison tool is global, so that literally any target audience can directly be reached. Results of TEDSrate-based artifact evaluations and comparisons become available instantaneously, which provides a great benefit also to developers if TEDSrate is used in pilot testing and iterative development cycles. The tests proved that time was little, cost was low, and resources were few that were needed to produce detailed artifact evaluations and real-world feedback.

These results give us confidence for asserting that TEDSrate has successfully addressed a core issue when it comes to improving and enabling timely and effective artifact evaluation.

Conclusion And Future Work

Software engineering success hinges on a number of variables, not all of which developers and software engineers are able to directly influence. However, many of those that can be directly addressed have also gone unattended for reasons of high cost, long time to complete, and prohibitive resource commitments necessary for producing meaningful and detailed feedback on artifacts. With the introduction of TEDSrate a tool has been created and tested that overcomes the cost, time, and resource barrier. It helps collect, analyze, and present detailed feedback data, which can immediately be used to adjust and change designs and improve artifacts.

In the next version of TEDSrate we will implement a post-processor, which transfers the numerical data to statistics packages for appropriate automatic analyses. We also consider the transfer of comments to an automatic text-mining post-processor.

Acknowledgments

Our thanks go to then graduate assistant Gary Gao, the developer of the alpha version of TEDSrate in 2013, who was followed by graduate students Delong Gao and Donghe Xu who added the initial version of the Admin Utility in 2014. Since 2015 William Menten-Weil has performed the technical design and development. We would also like to thank Janet Boyd, head of the graduate assistants crew at the Information School in the University of Washington, who helped get this project off the ground.

References

- [1] Abdelzad, V., T. C. Lethbridge, and M. Hosseini, "The role of semiotic engineering in software engineering," presented at the 5th International Workshop on Theory-Oriented Software Engineering (TOSE'16), Austin, TX, 2016.
- [2] Anvari, F., D. Richards, M. Hitchens, and M. A. Babar, "Effectiveness of persona with personality traits on conceptual design," presented at the ICSE'15, Florence, Italy, 2015.
- [3] Begel, A. and C. Sadowski, "2nd International workshop on user evaluations for software engineering researchers (USER 2013)," presented at the USER 2013 / ICSE'13, San Francisco, CA, 2013.
- [4] Buse, R. P., C. Sadowski, and W. Weimer, "Benefits and barriers of user evaluation in software engineering research," ACM SIGPLAN Notices, vol. 46, pp. 643-656, 2011.

[5] Chen, N., J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang, "AR-miner: mining informative reviews for developers from mobile app marketplace," presented at the ICSE'14, Hyderabad, India, 2014.

[6] Dawson, D., R. Desmarais, H. M. Kienle, and H. A. Müller, "Monitoring in adaptive systems using reflection," presented at the SEAMS, Leipzig, Germany, 2008.

[7] DeLone, W. H. and E. R. McLean, "Information systems success: The quest for the dependent variable," *Information & Management*, vol. 3, pp. 60-95, 1992.

[8] DeLone, W. H. and E. R. McLean, "The DeLone and McLean model of information systems success: a ten-year update," *Journal of management information systems*, vol. 19, pp. 9-30, 2003.

[9] Hurtado, N., M. Ruiz, E. Orta, and J. Torres, "Using simulation to aid decision making in managing the usability evaluation process," *Information and Software Technology*, vol. 57, pp. 509-526, 2015.

[10] Jurisch, M., H. Krmar, H. J. Scholl, K. Wang, Y. Wang, G. Woods, et al., "Digital and Social Media in Pro Sports: Analysis of the 2013 UEFA Top Four," presented at the 47th Hawaii International Conference on System Sciences (HICSS-47), Waikoloa, HI, 2014.

[11] Kitchenham, B., S. Linkman, and S. Linkman, "Experiences of using an evaluation framework," *Information and Software Technology*, vol. 47, pp. 761-774, 2005.

[12] Kolnay, P., "Ranking MLS' most popular teams; New research reveals fascinating results," in *World Soccer Talk* vol. 2016, ed. Florida, USA: World Soccer Talk, 2015.

[13] Nakamichi, N., K. Shima, M. Sakai, and K.-i. Matsumoto, "Detecting low usability web pages using quantitative data of users' behavior," presented at the ICSE'06. 28th International Conference on Software Engineering, Shanghai, China, 2006.

[14] Oh, J., S. Lee, and U. Lee, "How to Report App Feedback?: Analyzing Feedback Reporting Behavior," presented at the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, 2016.

[15] Ralph, P. and P. Kelly, "The dimensions of software engineering success," presented at the ICSE'14, Hyderabad, India, 2014.

[16] SBI. (2016, Sep 3). The Growth of MLS—Seattle Sounders Case Study [webpage]. Available: <http://sbibarcelona.com/the-growth-of-mls-seattle-sounders-case-study/>

[17] Schneider, K., S. Gärtner, T. Wehrmaker, and B. Brügge, "Recommendations as learning: From discrepancies to software improvement," presented

at the RSSE 2012. 3rd International Workshop on Recommendation Systems for Software Engineering, Zurich, Switzerland, 2012.

[18] Schneidewind, L., S. Hörold, C. Mayas, H. Krömker, S. Falke, and T. Pucklitsch, "How personas support requirements engineering," presented at the UsARE 2012. 1st International Workshop on Usability and Accessibility Focused Requirements Engineering, Zurich, Switzerland, 2012.

[19] Scholl, H. J., "Evaluating Sports Websites from an Information Management Perspective," in *Routledge Handbook of Sport Communication*, P. M. Pedersen, Ed., ed New York: Routledge, 2013, pp. 289-299.

[20] Scholl, H. J. and T. S. Carlson, "Professional Sports Teams on the Web: A Comparative Study Employing the Information Management Perspective," *European Sport Management Quarterly*, vol. 12, pp. 137-160, 2012.

[21] Scholl, H. J., M. Eisenberg, L. Dirks, and T. S. Carlson, "The TEDS Framework for Assessing Information Systems From a Human Actors' Perspective: Extending and Repurposing Taylor's Value-Added Model," *Journal of the American Society for Information Science and Technology*, vol. 62, pp. 789–804, 2011.

[22] Scholl, H. J., K. Wang, Y. Wang, G. Woods, D. Xu, Y. Yao, et al., "Top soccer teams in cyberspace: Online channels for services, communications, research, and sales," *Journal of Marketing Analytics*, vol. 2, pp. 98-119, 2014.

[23] Seddon, P. B., "A respecification and extension of the DeLone and McLean model of IS success," *Information Systems Research*, vol. 8, pp. 240-252, 1997.

[24] Shekhovtsov, V. A., H. C. Mayr, and C. Kop, "Stakeholder involvement into quality definition and evaluation for service-oriented systems," presented at the USER 2012. 1st International Workshop on User Evaluation for Software Engineering Researchers, Zurich, Switzerland, 2012.

[25] Taylor, R. S., "Value-added processes in the information life cycle," *Journal of the American Society of Information Science*, vol. 33, pp. 341-346, 1982.

