Performance Assessments of Android™-powered Military Applications Operating on Tactical Handheld Devices

Brian A. Weiss*, Lisa Fronczek, Emile Morse, Zeid Kootbally, Craig Schlenoff
National Institute of Standards and Technology, 100 Bureau Drive, MS8230,
Gaithersburg, MD, USA 20899-8230

ABSTRACT

Transformative Apps (TransApps) is a Defense Advanced Research Projects Agency (DARPA) funded program whose goal is to develop a range of militarily-relevant software applications ("apps") to enhance the operational-effectiveness of military personnel on (and off) the battlefield. TransApps is also developing a military apps marketplace to facilitate rapid development and dissemination of applications to address user needs by connecting engaged communities of end-users with development groups. The National Institute of Standards and Technology's (NIST) role in the TransApps program is to design and implement evaluation procedures to assess the performance of: 1) the various software applications, 2) software-hardware interactions, and 3) the supporting online application marketplace. Specifically, NIST is responsible for evaluating 50+ tactically-relevant applications operating on numerous Android™-powered platforms. NIST efforts include functional regression testing and quantitative performance testing. This paper discusses the evaluation methodologies used to assess the performance of three key program elements: 1) handheld-based applications and their integration with various hardware platforms, 2) client-based applications, and 3) network technologies operating on both the handheld and client systems along with their integration into the application marketplace. Handheld-based applications are assessed using a combination of utility and usability-based checklists and quantitative performance tests. Client-based applications are assessed to replicate current overseas disconnected operations (i.e., no network connectivity between handhelds) and to assess connected operations envisioned for later use. Finally, networked applications are assessed on handhelds to establish baselines of performance for when connectivity will be common usage.

Keywords: Performance Evaluation, Metrics, Android, Military, Functional Testing, Regression Testing, Usability

1 INTRODUCTION

The Transformative Apps (TransApps)[1] effort is a Defense Advanced Research Projects Agency (DARPA†)-funded program that began in 2010 and is aimed at enhancing the warfighter's effectiveness on and off the battlefield. Specifically, the program is developing a flexible and secure suite of applications ("apps"), enabling direct end-user input, promoting quick fielding and updates, and leveraging pre-existing state-of-the-art, commercial-off-the-shelf technology. To accomplish these goals, TransApps has focused its attention on developing a secure Android™‡ software platform, specialized application software, middleware and tools, a usable application portal, and flexible development processes. The program has made numerous achievements through January 2013, including developing over 50 apps for tactical users and fielding over 3000 handheld devices to warfighters in Afghanistan. Likewise, the program provided both first response and law enforcement personnel with over 150 devices to capture and share information in real-time to support the 2013 Presidential Inauguration. The program has received overwhelmingly positive feedback from the warfighter community including leadership personnel. Soldiers have credited the device with not only enhancing their

* Brian.weiss@nist.gov; phone 1 (301) 975-4373; fax 1 (301) 990-9688; www.nist.gov/el/isd

† This work was sponsored by the Defense Advanced Research Projects Agency (DARPA). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. . DARPA has reviewed this article and approved the content for public release and unlimited distribution.

‡ Certain commercial companies, products, and software are identified in this article to explain our research. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the companies, products, and software identified are necessarily the best available for the purpose.

situational awareness, but also enabling them to be successful in dangerous situations. The program continues to actively field additional units in Afghanistan in addition to providing application updates to existing users.

Testing is a critical element of this program. Personnel from the National Institute of Standards and Technology (NIST) have been funded to serve as an independent evaluation team since early 2011 for the TransApps program. NIST has been responsible for assessing three key areas of the TransApps program:

- Handheld Applications
- Client-based Applications
- Application Marketplace

The NIST evaluation team has extensive experience assessing advanced and emerging technologies; related prior work is discussed in Section 2. Greater detail on the TransApps program including an overview of the specific technologies tested can be found in Section 3. NIST's assessment of the handheld (HH) applications is discussed in Section 4; NIST's evaluation of client-based applications is presented in Section 5; and NIST's evaluation of the application Marketplace is discussed in Section 6. Finally, the conclusion is presented and future work is discussed.

2 NIST ADVANCED TECHNOLOGY ASSESSMENT EFFORTS

Personnel from NIST's Engineering Laboratory (EL) and Information Technology Laboratory (ITL) have extensive experience in evaluating advanced and emerging technologies for the military. NIST led the independent evaluation teams in assessing the DARPA Advanced Soldier Sensor Information System and Technology (ASSIST) technologies from 2004 to 2008, and the DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) technologies from 2006 to 2010. NIST's evaluation success is attributed to developing core competencies in designing and implementing evaluations, in addition to data analysis methods. Specifically, NIST developed the System, Component, and Operationally-Relevant Evaluation (SCORE) framework as the backbone of its test plan design and implementation for the ASSIST and TRANSTAC technologies. The following subsections will briefly discuss the ASSIST and TRANSTAC programs along with the SCORE framework.

2.1 ASSIST

Soldiers are often asked to perform missions that can take many hours. Examples of missions include presence patrols, search and reconnaissance, and apprehending suspected insurgents. After a mission is complete, the soldiers are typically asked to provide a report to their commanding officer describing the most important things that happened during the mission. This report is used to gather intelligence about the environment to allow for more informed planning for future missions. Soldiers usually provide this report based solely on their memory, still pictures, handwritten notes, or grid coordinates that were collected during the mission, provided these tools are available to the soldier. These missions are often very stressful for the soldier and there are undoubtedly many instances in which important information is not made available in the report, and is thus not available for the planning of future missions.

The ASSIST[2] program is a DARPA-funded effort that addressed this challenge by instrumenting soldiers with sensors that they can wear directly on their uniforms. The sensors include still cameras, video cameras, Global Positioning Systems (GPS), Inertial Navigation Systems (INS), microphones, and accelerometers. These sensors continuously record what is going on around the soldier while on a mission. When soldiers return from their mission, the sensor data is run through a series of software systems that index the data and create an electronic chronicle of the events that happened throughout the time that the ASSIST system was recording (see Figure 1) The electronic record includes times that certain sounds or keywords were heard, the times when certain types of objects were seen, and times that the soldiers were in a specific location or performing certain actions.

Figure 1: User Interface of Compiled Sensor Data

With this information, a soldier can give a report without relying solely on his memory. The electronic record will help jog the soldier's memory on activities that happened that he or she may not have recalled during the reporting period, or possibly even make him aware of an important activity that he or she did not notice when out on the mission. In addition, the multimedia information that is available in the electronic chronicle is available to the soldier to include in the report, providing substantially more information to the recipient of the report than text alone.

NIST's role in this program was to develop and implement evaluation procedures to characterize the performance of the software and technical components developed by external personnel. NIST focused on the evaluation of algorithms, software, and tools used to combine the passive data sensors and the automated event and object recognition capabilities.[3,4]

Specific technologies developed for ASSIST and evaluated by NIST include:

- Object Detection / Image Classification – the ability to recognize and identify objects in the environment,
- Arabic Text Translation – the ability to detect, recognize, and translate written Arabic text,
- Sound Recognition / Speech Recognition – the ability to identify sound events (e.g., explosions, gunshots, and vehicles) and recognize speech,
- Shooter Localization / Shooter Classification – the ability to identify gunshots in the environment,
- Soldier State Identification / Soldier Localization – the ability to identify a soldier's path of movement around an environment, and to characterize the actions taken by the soldier.

2.2 TRANSTAC

Clear and effective communication is crucial to developing peaceful and productive relationships with foreign populations, especially when U.S. servicemen and women are operating on foreign soil. Military personnel operating in foreign countries are often confronted with communication challenges. Successful communication has historically relied upon the availability of competent and cooperative interpreters and/or military personnel that are bilingual. The U.S. military identified that the demand for reliable and qualified interpreters was greater than their availability. DARPA sought to solve this problem and began to explore speech translation technologies thereby developing the TRANSTAC program. The goal of the TRANSTAC program[5] was to demonstrate capabilities to develop and rapidly field free-form, two-way translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter.

Several prototype systems were developed under this program for numerous military applications, including force protection and medical screening. The technology has been demonstrated on smartphone (shown in Figure 2) and laptop platforms. NIST was asked to assess the usability of the overall translation system and to assess each individual component of the system (specifically, automatic speech recognition, machine translation, and text-to-speech) [6,7].

Figure 2: TRANSTAC System on a Smartphone Platform

All of the TRANSTAC systems employ a similar workflow. Either live English speech or a recorded audio file is fed into the system. Automatic Speech Recognition processes the speech to recognize what was said and generates a text file of the speech. That text file is then translated to another language using Machine Translation technology. The resulting text file is then spoken to the foreign language speaker using Text-To-Speech technology. This same process then happens in reverse when the foreign language speaker speaks. These two-way, real-time, voice-translation devices are designed to improve communications between the U.S. military and non-English speakers in foreign countries.

2.3 SCORE

NIST developed the SCORE framework; SCORE is a unified set of criteria and software tools for defining a performance evaluation approach for complex intelligent systems. It provides an evaluation framework that assesses the technical performance of a system and its components through isolating and changing variables, and capturing end-user utility of the system in realistic use-case environments. SCORE is built around the premise that, in order to get a comprehensive picture of how a system performs in its actual use-case environment, technical performance should be evaluated at the component and system levels.[8,9,10]

The SCORE framework advocates identifying evaluation goals and user requirements, and then identifying evaluation methodologies that support those test parameters. Once the set of evaluation methodologies have been identified that can support the evaluation, then method selection can be further refined by other logistical parameters. These parameters could include the availability of qualified personnel to design and conduct the assessment, the type of testing environment that is needed to execute the test, the mechanisms that are needed to collect the data, and the data analysis considerations (e.g., whether the time and resources exist to code many hours of video data).

SCORE takes a tiered approach to measuring the performance of intelligent systems. At the lowest level, SCORE uses elemental tests to isolate specific components, and then systematically modifies variables that could affect the performance of that component to determine those variables' impact. Typically, elemental tests are performed for each relevant component of the system. At the next level, the overall system is tested in a highly structured environment to understand how modifying specific variables impacts the overall system performance. In the context of speech translation, example variables that would be modified to assess their impact on performance would be background noise, speaker dialect, and speaker gender. Next, individual capabilities of the system are isolated and tested for both their technical performance and their utility using task tests. Lastly, the technology is immersed in a longer scenario that evokes typical situations and surroundings in which the end-user is asked to perform an overall mission or procedure in a highly-relevant environment which stresses the overall system's capabilities. Formal surveys and semi-structured interviews are used to assess the usefulness of the technology to the end-user.

SCORE is unique in that:

- it is applicable to a wide range of technologies, from manufacturing to defense systems,
- elements of SCORE can be decoupled and customized based upon evaluation goals,
- it has the ability to evaluate a technology at various stages of development, from conceptual to full maturation, and

- it combines the results of targeted evaluations to produce an extensive picture of a system's capabilities and utility.

3 TRANSFORMATIVE APPS

DARPA selected NIST to be a primary evaluator of the TransApps technologies in 2011. Specifically, NIST was tasked to assess the performance of the 1) experimental handheld applications, 2) client-based applications, and 3) on-line application Marketplace. As an independent, third-party evaluation team, NIST personnel presented unbiased and objective performance data to the DARPA sponsor, enabling them to make informed decisions regarding application stability and field-worthiness. This multi-disciplinary evaluation team, comprised of engineers and computer scientists from NIST's Engineering and Information Technology Laboratories, facilitated frequent interactions among other testers and developers to deeply understand the scope and intent of the applications. Likewise, the NIST team embraced the importance of end users' voices and frequently sought out those representing the end user population.

The NIST team leveraged the principles of SCORE to develop and implement test protocols and procedures to yield comprehensive performance assessments at multiple layers. This includes individual assessments of apps while isolated from other apps, and global assessments of the apps and their interactions while operating on configurations expected during fielding. Likewise, protocols also enabled NIST personnel to assess the technology in tightly controlled indoor environments and less-controlled outdoor environments. The protocols developed and employed by NIST in this effort include:

- Key function testing – Specifically-developed protocols enable NIST testers to conduct an immediate assessment of an app's performance in key functions yielding data at a high-level.
- Functional regression testing – Aimed at verifying the performance of pre-existing functions and features as other functions and features are added or updated. This testing produces detailed data and findings.
- Usability feedback testing – Provides qualitative assessments based upon NIST usability expertise in mobile and advanced technologies.

These protocols are discussed throughout Sections 4, 5, and 6 as they relate to the three key areas of NIST assessment focus; 1) Handheld application assessment, 2) Client-based applications, and 3) Application Marketplace.

4 HANDHELD APPLICATION ASSESSMENT

Handheld devices that are being deployed to overseas U.S. military personnel are provisioned with over 50 apps. Many of the apps perform a single function (e.g., one app displays the device's current altitude and another displays the Julian date). Other apps provide a set of related functions (e.g., one app performs a host of common unit conversions). Several apps serve as training instruments. For example, in the language domain, apps like Afghan Phrases and MilTrans help the user learn rudimentary communication skills; and there are many apps devoted to teaching medical and survival skills. Finally, there are a few apps that provide core support to the war-fighter; these "core" apps typically integrate several apps into a single interface. Some examples of the core apps are presented in Section 4.1, while the NIST team's testing methodology and protocols are presented in Section 4.2.

4.1 Examples of Core Apps

The app called Maps is intended to be the primary focus of soldiers when they are on a mission. The app's background is a satellite imagery map. Since all of the devices are GPS-enabled, it is possible for Maps to render the current location of the user on the map. Objects that are aware of geo-location can be rendered on the map as layers. Examples include GPS traces of user location information, user-generated annotation markers called Spots, markers for media collections, and specialized markers for other apps that contain geo-location info.

PLI (Personal Location Information) is a core app that integrates with Maps and is responsible for recording the location of the handheld device. At the interface, this app allows the user to turn recording off and on and to turn broadcasting of position information off and on. This enables the user to view their own position and share this information with other users. This data is very important to the user, so the information is available not only as a desktop widget, but is shown in the status bar at the top of the screen. Figure 3 presents a screenshot of the Maps app with PLI data.

Figure 3: Primary maps application showing the user's own location (green triangle), surrounding area, and PLI (blue circles) of other team members

TRAQ (Tactical Recording and Acquisition) is a core app that provides the user with the ability to collect audio, video, and still pictures that are geo- and time-stamped. In addition to real-time collection support, the TRAQ app has a tool called Review. Review takes the trace data collected by PLI and the geo- and time-stamped media collected with the TRAQ recorders, and renders the view as a timeline of activity superimposed on a map. One key use of Review is to allow the user to select and export segments of the timeline so that "heat maps" can be generated. Figure 4 presents a screenshot of TRAQ.

Figure 4: TRAQ (Tactical Recording and Acquisition) provides users with geo-referenced media such as picture, video, and audio options and a review of historical traces with the associated geo-located media

Heatmaps (see Figure 5) are used in two ways on the current handhelds. As a single user tool, Heatmaps interacts with Maps to allow the user to select one or more "sessions" (i.e., time segment) to produce a color-coded view of where the user traveled, the frequency of location visitation, and the duration of visits. The resultant "heat map" is a user-selectable layer within the Maps app. Regions that are coded red are places that the user either visited often or stayed a long time. Blue indicates regions with minimal presence. The rest of the spectrum indicates intermediate use patterns. The second way that heat maps are used is to combine traces exported from TRAQ from several users. These maps show where those different people were located over some time period. According to the users of heat maps, the artifacts are useful to identify gaps in coverage (e.g., where in this town has the squad not been?) and to identify places that are too highly utilized (e.g., routes that the enemy could target because they are preferentially used).

Figure 5: Recorded GPS information rendered into a heat map visualization overlaid on Maps

The Maps app supports the concept of plugins. Plugins are defined as apps that run over-top one or more existing apps. For example, PLI and Heatmaps are plugins for Maps (i.e., Maps must be installed in order to use either of these two plugins). Two of the main plugins are Navigation Tool (also called Route Planning) and Bearing Tool. Both apps allow the user to create locations by tapping on the visible map or by entering geographical coordinates of points. The user of these apps can then adjust points by dragging them to a new location on the map or by entering different coordinates. These plugins are highlighted in Figure 6.

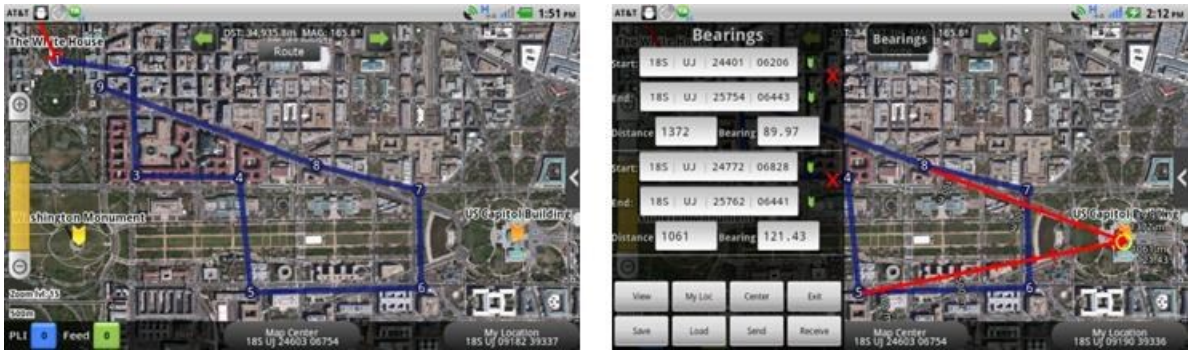


Figure 6: Route Planning (on the left) and Bearing (on the right) are examples of Map Plugins

4.2 Testing Practices for Handheld Applications

The above description of apps reveals only the tip of the iceberg when trying to develop a testing methodology. Apps range in behaviors. Some apps are simple and direct and others are complex and have many moving parts. Besides app complexity, the major factor in determining what type of testing is needed often depends on whether the apps are "connected" or not. When a user is sent into the field and there are no communications channels enabled, this is said to be in "disconnected" mode. When communication channels such as tactical radios are available, then the devices are said to be "connected."

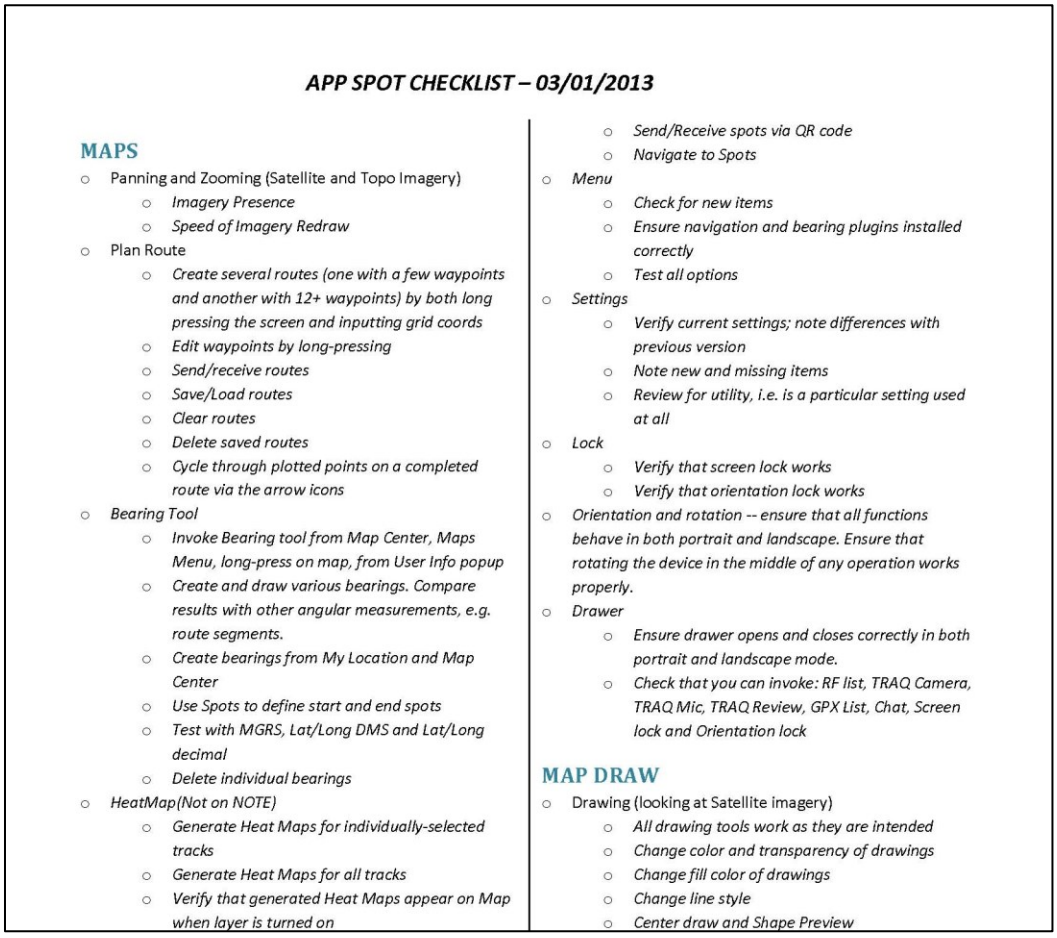
In disconnected mode, the typical workflow is that, before a mission begins, users are briefed face-to-face by their leaders. At that time the individuals are given devices that are pre-loaded with information that they may need to use on the mission. These pre-loads can included drawing overlays, heat maps, sets of spots, images of people and locations of interest, and planned routes. During a disconnected mission, individual users move throughout the physical environment and collect information. That information is stored locally and, upon return to base, the user can transfer the data to larger computers where the data can be stored and manipulated.

In connected mode, workflow is much more variable due to real-time communication. Communication apps include Chat that supports attachment of pictures, videos, audio, spots, new routes, drawings, and various other types of information. In addition, when devices are connected over a network, users can share their location information with

others. At the interface, connectedness produces a richer picture of the tactical situation. It is precisely the increase in richness that makes testing more difficult.

The NIST team uses a user-centered approach to testing handheld devices and apps. The fact that our target users (i.e., soldiers operating overseas) are not available for real-time testing has led NIST to develop alternative routes to incorporating user surrogates as part of its testing. Briefings with soldiers who recently returned from overseas deployments provide the development and testing teams with many insights about how the apps performed. These people are excellent sources of desired new features and brand-new apps. The program supports field-service representatives who act as technical and training liaisons with deployed units. Testers have access to these personnel on weekly teleconferences and via their written materials. Using these sources, we have developed an understanding of the typical workflows of users. The NIST team has leveraged its understanding of user workflows to develop relevant app use cases.

Expert and heuristic reviews[11] are the primary day-to-day methods that the NIST team uses to evaluate apps. For each app we have enumerated the user interface (UI) components and the functions that each app is intended to support. We refer to this as the App Spot Checklist (shown in Figure 7). We find that this simple tool can be used by novice users to guide their initial explorations, and seasoned testers use it during regression testing. Ultimately, the test team communicates findings in two distinctly different ways. First, weekly reports are delivered to upper level program management, and in these reports the show-stoppers and watchpoints are itemized and prioritized. Second, bugs, suggestions for new features, and other ideas are submitted to an online bug reporting system tool. The system is used primarily by developers and managers to track progress. The workflow of the typical bug life-cycle includes a testing step that forces all issues to be viewed by testers before the bug is addressed.



APP SPOT CHECKLIST – 03/01/2013

MAPS

- Panning and Zooming (Satellite and Topo Imagery)
  - *Imagery Presence*
  - *Speed of Imagery Redraw*
- Plan Route
  - *Create several routes (one with a few waypoints and another with 12+ waypoints) by both long pressing the screen and inputting grid coords*
  - *Edit waypoints by long-pressing*
  - *Send/receive routes*
  - *Save/Load routes*
  - *Clear routes*
  - *Delete saved routes*
  - *Cycle through plotted points on a completed route via the arrow icons*
- *Bearing Tool*
  - *Invoke Bearing tool from Map Center, Maps Menu, long-press on map, from User Info popup*
  - *Create and draw various bearings. Compare results with other angular measurements, e.g. route segments.*
  - *Create bearings from My Location and Map Center*
  - *Use Spots to define start and end spots*
  - *Test with MGRS, Lat/Long DMS and Lat/Long decimal*
  - *Delete individual bearings*
- *HeatMap(Not on NOTE)*
  - *Generate Heat Maps for individually-selected tracks*
  - *Generate Heat Maps for all tracks*
  - *Verify that generated Heat Maps appear on Map when layer is turned on*
  - *Send/Receive spots via QR code*
  - *Navigate to Spots*
- *Menu*
  - *Check for new items*
  - *Ensure navigation and bearing plugins installed correctly*
  - *Test all options*
- *Settings*
  - *Verify current settings; note differences with previous version*
  - *Note new and missing items*
  - *Review for utility, i.e. is a particular setting used at all*
- *Lock*
  - *Verify that screen lock works*
  - *Verify that orientation lock works*
- *Orientation and rotation -- ensure that all functions behave in both portrait and landscape. Ensure that rotating the device in the middle of any operation works properly.*
- *Drawer*
  - *Ensure drawer opens and closes correctly in both portrait and landscape mode.*
  - *Check that you can invoke: RF list, TRAQ Camera, TRAQ Mic, TRAQ Review, GPX List, Chat, Screen lock and Orientation lock*

MAP DRAW

- Drawing (looking at Satellite imagery)
  - *All drawing tools work as they are intended*
  - *Change color and transparency of drawings*
  - *Change fill color of drawings*
  - *Change line style*
  - *Center draw and Shape Preview*

Figure 7: Excerpt from the NIST App Spot Checklist

5 CLIENT-BASED APPLICATIONS

Client-based applications consist of tools developed for encrypted laptops (clients) and are of paramount importance when network connectivity is limited or absent. Client-based applications are used during pre- and post-mission practices for data creation and for data transfer between clients and handhelds. Section 5.1 gives a brief description of the main client-based applications, while Section 5.2 describes testing practices that the NIST team has created to evaluate the qualitative and quantitative performance of client-based applications.

5.1 Overview of Client-based Applications

Data downloads and uploads between handhelds and clients are achieved with Device Manager. Device Manager allows information to be uploaded to the handhelds for the next mission and to download data from the handhelds (see Figure 8) to the client after a mission has been carried out. The scenarios commonly used in the field are described as follows:

- Before a mission, the user sets up each handheld with essential information that each unit will need to target during duty in the field. An example of such information is a map of the area with drawings representing sensitive buildings, the route to take to reach different critical areas, and locations of specific meetings. Once all necessary files are ready, the data is selected and uploaded to the handheld devices.
- When a unit returns from its mission, the data captured in the field are transferred to the clients. The collected data are then compiled in different formats for different purposes. For instance, data retrieved from a unit's handheld can be used to display the path taken by the unit along with pictures, videos, and drawings captured during the mission. All these materials can be displayed on the same map for deep tactical analysis.

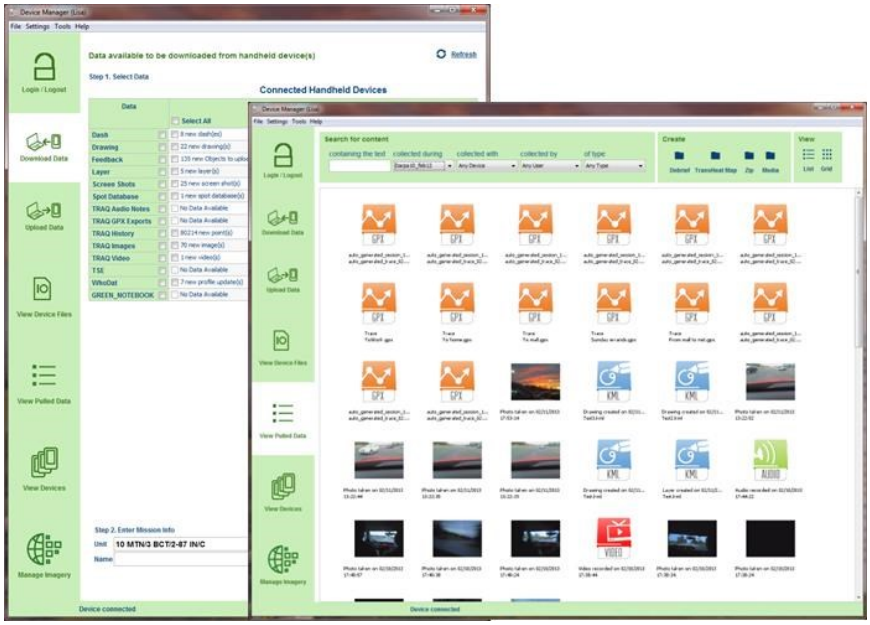


Figure 8: Device Manager simplifies communication between from handheld to PC (left) and from PC to handheld

In addition to data downloads and uploads, a key component of Device Manager is its ability to combine downloaded data to generate different file formats (e.g., zip files, Powerpoint presentations, heat map files). The heat map file generated by Device Manager is readable by the handheld application called TransHeat. TransHeat displays a heat map of which locations have seen the most activity (similar to the heat map presented in Figure 5). This application has been used to conduct analysis for targeting purposes. A heat map can be assimilated to patterns seen through the eyes of insurgents; therefore, leaders also refer to the displayed heat map when planning future missions so as to not repeat existing patterns. In missions where multiple patrol leaders share the same space, a heat map gives information on areas to avoid where the patrol would have otherwise not recognized as danger areas based on their peer's patrol history.

Users have another strategic tool called Tactical Maps (nicknamed BLOX – Figure 9) at their disposal to keep track of mission patrols when their client(s) are connected to a network. BLOX allows a user to track each unit member via their

PLI (PLI window, Figure 9; lower-right dialog window). The PLI window displays the list of unit members (*Name* column), the last time their status was updated to the BLOX server (*Last Updated* column), and their current connectivity (blue circles when connected to BLOX, and gray circles when disconnected). When radio connectivity is available (i.e., handheld users in the field have network capability), unit patrols can share information with other unit patrols (handhelds to handhelds), with client users (handhelds to BLOX), or both. BLOX also allows client users to send updated information to the units in the field via text messages, pictures, videos, and drawings (*Feed* window, Figure 9, upper-right dialog window). This feature is beneficial when new orders are received to update the current mission (e.g., changed routes, or order to search a new area).

BLOX is also a powerful tool to create drawings through the *Map Drawing* feature. The *Map Drawing* window (Figure 9; left-most dialog window) shows a great variety of tools ranging from simple lines to polygons and grids. Numbers and texts can be added to the drawings for more accuracy during tactical operations. This drawing function can be useful in either "connected" or "disconnected" mode. In "connected" mode, drawings can be sent directly to handhelds in the field and in "disconnect" mode, they can be exported to KML (Keyhole Markup Language) files to be transferred to handhelds via Device Manager.

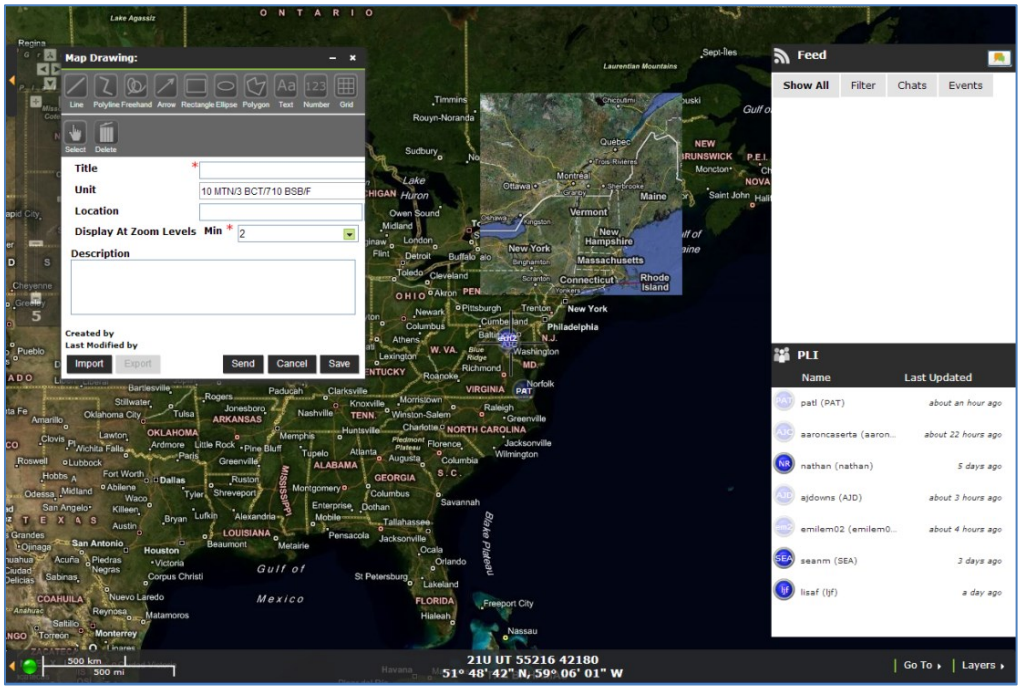


Figure 9: User interface of the BLOX web-based application.

Noting the principal role of client-based applications in TransApps has led the NIST team to develop functional regression testing presented in the following section.

5.2 Testing Practices for Client-based Applications

The NIST team needs to assure that common practice scenarios are repeatable and reproducible to assess the performance evaluation of client-based applications. Therefore, it is important to simulate the roles of the various military personnel expected to use the client applications. Moreover, carrying out the same tasks under the same conditions observed in the field will provide more realistic results. The rest of this section describes each procedure performed on the client and handheld sides along with the techniques used to evaluate the performance of these procedures.

- Device Setup – Connecting a handheld to a client requires a series of steps (provided by the program developers) that assure a secure and successful communication between both devices, because clients and handhelds are encrypted and will carry sensitive information when fielded. The NIST team's first approach is to omit all the steps and check that the client is not interacting with the handheld to make sure that all the required

steps are necessary to lead to the expected behavior. If the result is successful, i.e., the client does not communicate with the handheld, the second required step is added from the list in our testing method, and we start testing from the first step. This method is pursued until the first occurrence of a failed or successful result. This testing method confirms that all the specified steps are mandatory for the client and the handheld to communicate with each other.

- Data Upload – Unit patrols in the field carry handheld devices on which the latest information for the current mission has been uploaded. Uploaded data are of different natures and of different sizes. It is important to test that the files selected in Device Manager to be transferred to a handheld really appear on the handheld and are consistent with the original files. To tackle the aforementioned aspects, the NIST team first generates very complex files on the client. For instance, a very complex drawing can be generated with the use of each single tool available in the *Map Drawing* library (Figure 9). The purpose of this testing is to stress the limit of data rendering and data transfer from a client to a handheld. It is common to see complex drawings not properly rendered on a handheld due to loss of information during the transfer, hardware or software limitations, and applications issues. Pushing the boundaries during testing helps to identify issues at an early stage before deploying devices and applications in the field.
- Data Collection – A unit investigates a specific area and captures relevant information in the form of pictures, videos, audios, and drawings during a mission. Although no performance evaluation is performed on the client during data collection, it is important to set the handhelds properly to avoid future data loss. The NIST team measures the performance of handheld-based applications during the setup and during the data collection routine. For instance, assume the tester notes the time it takes to receive a GPS signal after turning on the handheld for the first time versus a handheld already turned on. The tester compares the real time synchronization frequency with the frequency that has been set up in the settings menu. During data collection, the tester quantitatively measures the time the handheld takes to capture and save pictures, videos, and audios. With the GPS receiver activated, the tester also makes sure that the traces built in real time match the current path that the tester is following. The testing approach adopted for data collection is quite different from the ones used for setting up devices and data uploads since data collection is done solely with the handhelds and does not involve client usage.
- Data Download – At the end of a mission the data is retrieved that was collected during the time the user spent in the field. The handhelds are plugged into the clients and downloaded via Device Manager. There is currently no automatic method to detect discrepancies between the data that have been collected (by a unit patrol) in the field and the ones downloaded on the client since the current field procedures call for another individual (not the handheld user) to download the data. To qualitatively test this scenario, the NIST testers involved in the field exercise are asked to validate the accuracy of the data they have collected. For instance, if a heat map produced from data collected by tester A indicates a hot spot for an area that tester A never visited, tester A will be able to invalidate this fact. This information is captured in the form of additional notes so that no data is lost in the reporting process. In the same way, if a picture has been taken at location 1 by tester A and shows up on the client at location 2, tester A will also be able to invalidate this fact. Identifying these types of problems during testing anticipates inaccurate data analysis in the post-mission meetings in the field.

6 APPLICATION MARKETPLACE

6.1 Overview of Marketplace

Marketplace, a web-based military app store, serves the needs of a broad range of users (military and civilian), with each user having a particular job role and skill level. It is a one-stop shop for users, app developers, and support personnel to get the latest TransApps applications for their handhelds, offer feedback on application successes and shortcomings, learn more about the program, share ideas, access tutorials and message boards, and vote on ideas to improve apps. Marketplace provides a mechanism to update handhelds rapidly and effectively, and encourages interactions between users, developers, and support personnel.

Marketplace is a one-stop shop that is fully web-based. It has an app catalog for the general user (see Figure 10), an imagery warehouse to meet the needs of users in different parts of the world, a Developer dashboard to support the current and future development team, and an Admin dashboard to support program management and system

administrators. Access to Marketplace is by invitation only from program administrators. Users logon to the site and gain access based on tiered permission levels.

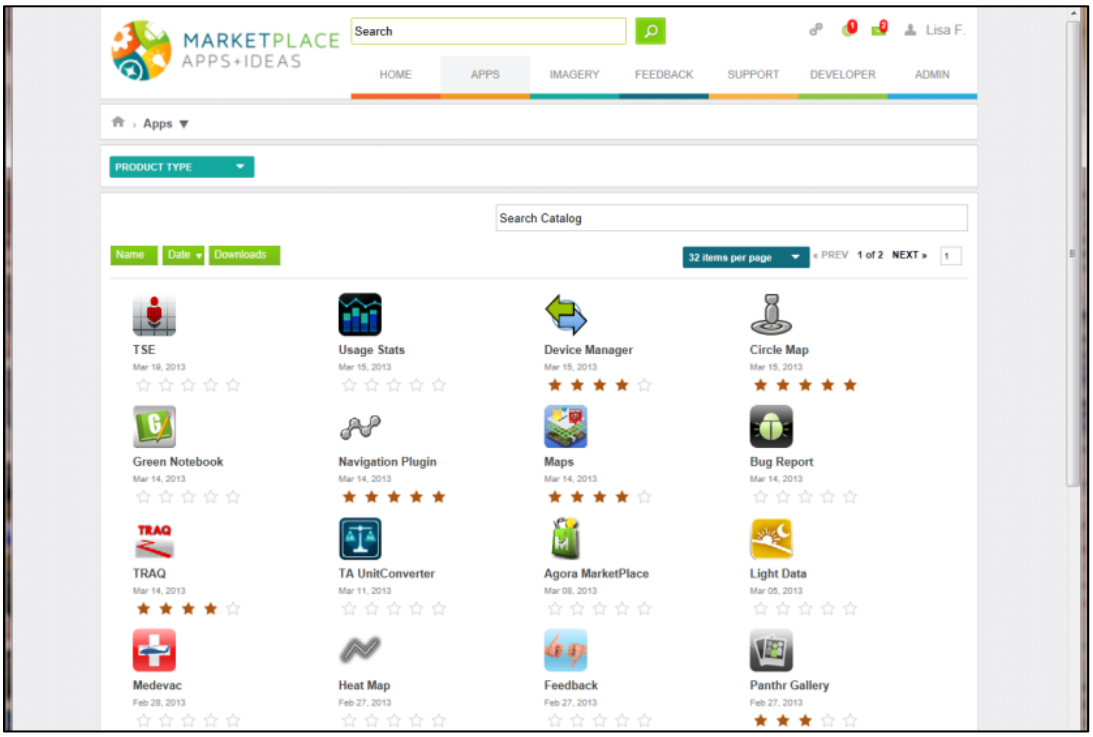


Figure 10: Marketplace app catalog screenshot

Marketplace enables various personnel the freedom to access the features needed for their specific activities. The typical end-user may only be able to see the App Catalog, to read and write reviews, feedback, and ideas, to view training videos, to access message boards, to vote on ideas, and to access help documents. Third-party app developers may see the App Catalog and the status of any apps that they have loaded into the site, but not the other apps that have been uploaded into the system but not yet approved for general use. Approved app developers (i.e., developers already known to the program) and TransApps development team have access to the full app catalog, and the full set of apps that are in the "developer" space, and software for use on the client computer system (that is, the computer that is used to support handhelds in the field) and changes needed for kernel modification. Marketplace system administrators and program managers have access to the full site.

Users can browse the app catalog, write reviews, and learn more about the TransApps program, apps, and uses for the handheld. Users earn points and challenge coins for regular participation in the site's functions with acts as simple as viewing the current app catalog, providing feedback on the program and apps, and even replying to others comments, suggestions, or questions. Military leaders have been using challenge coins for years to increase morale, encourage a sense of belonging, and reward acts worthy of recognition. The use of challenge coins gained popularity even outside the immediate military environment since it encourages friendly competition. Marketplace built upon this idea as way to encourage frequent interaction of its user community. Marketplace makes use of social media tools such as voting (on apps, ideas, feedback, etc.) to give its users an effective way to get their ideas heard and to quickly confirm whether this idea is valid.

Marketplace has personalized interfaces for its users. Regular users (civilian and military) may only see the Home, general information, and the App Catalog specific to their community needs. Developers have their own interface to upload handheld and client apps, kernel modifications, imagery, and help documents. The term "Developers" has a broad meaning. It refers to the research team that is developing the TransApps program, program managers, and the testing team. It also refers to app developers. Currently all of the app developers are pre-approved developers that focus on the most immediate needs of its users (either military or first responders). In the future, the TransApps program plans to open up the program to software entrepreneurs to develop apps for tactical battlefield, humanitarian, disaster recovery,

and other missions. App functions might include command and control, reporting, mission planning, intelligence/surveillance/reconnaissance, real-time collaboration, geospatial visualization, analysis, language translation, training, and logistics tracking. A process has been developed that allows app developers to upload their source code where it will be analyzed for security flaws and compiled. At that point, it will be placed in a holding area until it can be evaluated and published. "Kits" of specialized apps are geared to specific audiences (e.g., first responders, military users outside of the contiguous United Status, and those military users who are stateside). For example, users in the United States can take advantage of "connected" apps and wide availability of network options (mentioned above), while those outside of the contiguous United States often lack reliable network and use "disconnect" apps primarily.

6.2 Testing Practices for Marketplace

Marketplace needs to serve many user groups, making it complex to evaluate. The system must determine who the user is and what information she needs to see, access, or even modify. Should a tester try to "play" different roles (e.g., system admin, imagery person, developer, or basic user) and then try to perform those functions? Personas are useful in this type of testing since they create a common shared understanding of the user group – allowing the developer to build the design around this process. Also, the personas help to prioritize the design considerations by providing a context of what the user needs and what functions are nice to add and have.

The Marketplace development team originally built five personas into the system: Joe (a basic user with no specialized access), Joe Military (another basic user who has the ability to view military-specific information), Joe Kit Maker (a mid-level user who can make "kits" for specific audiences), Joe Test Developer (an app developer who has been pre-approved to supply apps for the TransApps program), and Joe Developer (an app developer with no prior affiliation to the program who believes they have a useful app for the project.). The "Joe" personas allow the testing team to quickly change their access levels. The testing team can probe the system to see which functions can be accessed and which cannot. As the project became more complex, it became clear that other personas needed to be tested, too (e.g., Program Manager, Help Center team).

As with any website there are many levels of testing that can happen. Does something work or not? Or even as simple as, does the website open and allow a user to log on? Are the hyperlinks valid? Do the hyperlinks navigate to the correct location? Is it easy to navigate? Can the user find the information for which he is looking? Is there more than one way to find that information? Are the navigation elements consistent from page to page? Does a page take a long time to load – longer than reasonable?

After the more direct tests are run, it is important to compare the site with its intended usage. For Marketplace, does it meet the needs of the various TransApp team members including its users? Does the site restrict information to the user's access tier? Are user roles maintained as the user navigates from one hyperlink to another? Does the site support the technical team that outfits the handhelds both here in the U.S. and abroad?

Website testing is often very subjective. For example, what is thought to be an easy-to-navigate page for one person may not be for another; however, "does the webpage open in Internet Explorer?" is an objective test. Open communication between the testing and development team was important to establish early on. It reinforced our common goal to provide the best package possible for both military and civilian users. While the goal for Marketplace has remained consistent to serve TransApps users, developers, and support teams, the user interface has changed dramatically from being a complex set of navigational links that floated around the screen to a bright easy-to-use site. Marketplace users now have unique dashboards that are intuitive, user-friendly, and flexible.

Functional regression testing is an integral part of testing the Marketplace given that the developers and testers have employed a weekly development and test cycle since 2011. In a given week, the developers will identify errors (both critical and trivial) to be fixed and additions to be made. The development cycle concludes with the testers verifying and validating these changes according to documentation provided by the developers. The test team provides detailed feedback to the development team indicating which features were still functioning as intended, what fixes were successfully in place, and what areas still required attention. Effective functional regression testing became a very dynamic process as Marketplace continued its evolution and increased in complexity.

7 CONCLUSIONS AND FUTURE WORK

The NIST evaluation team has made (and continues to make) a significant impact on the TransApps program in their extensive and detailed testing of the handheld applications, client applications, and Marketplace portal. Since the NIST

team began assessing applications for this program in 2011, NIST test feedback and reports have offered program leadership greater insight into the capabilities and limitations of the TransApps technologies. This has led to more informed and quicker fielding of the technology. Not only is the program leadership aware of the technology's latest performance, program personnel in Afghanistan are knowledgeable of the latest iterations of the technology before updating their devices.

NIST testers continue their daily interactions with program staff including other testers, developers, and leadership to ensure test efforts are focused on the highest priority apps and testing occurs at the appropriate levels. Test practices and protocols are frequently reviewed and updated to adapt to changing priorities and the inclusion (or removal) of specific applications. Functional regression testing, usability testing, and high level core functionality testing are regularly conducted across the program's technology as fixes are made and apps are upgraded. The backbone of the NIST team's effort continues to be its independent, third-party objectivity to offer the program sponsor unbiased and thoughtful feedback detailing technological capabilities and limitations. Moving forward, the NIST team's expertise is being leveraged to work with other program staff to develop targeted tests to assess the capabilities of various hardware platforms prior to their assimilation into the program. The NIST team expects the technology's evolution to continue to enable end-users to work safer and more efficiently in challenging and threatening environments.

8 ACKNOWLEDGEMENT

This work was supported by the Defense Advanced Research Projects Agency (DARPA) TransApps program led by the Program Manager, Doran Michels.

REFERENCES

[1] http://www.darpa.mil/Our_Work/I2O/Programs/Transformative_Apps.aspx

[2] Schlenoff, C.I., Steves, M.P., Weiss, B.A., Shneier, M.O., and Virts, A.M., "Applying SCORE to Field-Based Performance Evaluations of Soldier-Worn Sensor Technologies," Journal of Field Robotics – Special Issue on Quantitative Performance Evaluation of Robotic and Intelligent Systems, vol. 24, 671 – 698, (September 2007).

[3] Schlenoff, C.I., Weiss, B.A., Steves, M.P., Virts, A.M. and Shneier, A.M. "Overview of the First Advanced Technology Evaluations for ASSIST," Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop, (2006).

[4] Weiss, B.A., Schlenoff, C.I., Shneier, M.O., and Virts, A.M. "Technology Evaluations and Performance Metrics for Soldier-Worn Sensors for ASSIST," Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Worksho*p*, (2006).

[5] Schlenoff, C., Weiss, B., Steves, M., Sanders, G., Proctor, F., and Virts, A., "Evaluating Speech Translation Systems: Applying SCORE to TRANSTAC Technologies," Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference (2009).

[6] Weiss, B.A. and Schlenoff, C., "Performance Assessments of Two-way, Free-Form, Speech-to-Speech Translation Systems for Tactical Use," Proceedings of the 2010 International Test and Evaluation Annual (ITEA) Symposium, (2010).

[7] Weiss, B.A., Schlenoff, C., Sanders, G, Steves, G., Condon, S., Phillips, J., and Parvaz, D., "Performance Evaluation of Speech Translation Systems," Proceedings of the 6th edition of the Language Resources and Evaluation Conference (2008).

[8] Schlenoff, C., "Applying the Systems, Component and Operationally-Relevant Evaluations (SCORE) Framework to Evaluate Advanced Military Technologies," ITEA Journal of Test and Evaluation, 31(1) (February 2010)

[9] Weiss, B.A. and Schlenoff, C.I., "The Impact of Evaluation Scenario Development on the Quantitative Performance of Speech Translation Systems Prescribed by the SCORE Framework," Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop (2009).

[10] Weiss, B.A. and Schlenoff, C.I., "Evolution of the SCORE Framework to Enhance Field-Based Performance Evaluations of Emerging Technologies," in Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop (2008).

[11] Nielsen, J., Heuristic Evaluation in [Usability Inspection Methods], Nielsen, J. and Mack, R.L., editors, John Wiley & Sons, New York, NY (1994)