The development of a test harness for biometric data collection and validation

Michael E. Brockly, Stephen J. Elliott*, Robert W. Proctor
International Center for Biometric Research,
Purdue University,
West Lafayette, IN 47907 USA
*elliott@icbrpurdue.org

Richard Guest
School of Engineering and Digital Arts,
University of Kent
Canterbury, Kent, CT2 7NT, UK

***Abstract*—Biometric test reports are an important tool in the evaluation of biometric systems, and therefore the data entered into the system needs to be of the highest integrity. Data collection, especially across multiple modalities, can be a challenging experience for test administrators. They have to ensure that the data are collected properly, the test subjects are treated appropriately, and the test plan is followed. Tests become more complex as the number of sensors are increased, and therefore it becomes increasingly important that a test harness be developed to improve the accuracy of the data collection. This paper describes the development of a test harness for a complex multi-sensor, multi-visit data collection, and explains the processes for the development of such a harness. The applicability of such a software package for the broader biometric community is also considered.**

***Keywords*—Biometrics, data collection, human error, test administrator error, test harness**

I. Introduction (*Heading 1*)

Biometrics is defined as the automated recognition of individuals based on their behavioral and biological characteristics [1]. A number of factors constitute a biometric system, including the biometric characteristic itself, the sensor, the subject, the algorithm, the environment, and the test administrator. The test administrator is a critical part of a biometric data collection system, being responsible for following data collection procedures and supervising the test subjects [2]. The test administrator is also responsible for monitoring the data quality prior to the data being accepted into the data collection system. Data that are being monitored include biometric samples as well as metadata, such as age, ethnicity, etc. It is important that both types of data are correct when entered into the system.

Poor data quality refers to biometric data that are captured incorrectly, causing low image quality, incorrect labeling of biometrics, or incorrect entering of metadata. At the operational level, poor data, regardless of the source, lead directly to subject dissatisfaction, increased cost, and lowered test administrator job satisfaction [3].

Metadata also needs to be acquired correctly. Metadata are important in data collections because they provide additional context to the biometric samples. Examples include gender, moisture of subjects' fingerprints, and documentation of any disorders that may affect the subject's ability to complete the successful presentation of a biometric sample. The task of entering and updating biometric data into a database can create metadata errors [4].

Furthermore, these data are typically entered manually in the presence of the subject. If the metadata information are incorrect, the results of the data analysis will be inaccurate because subjects may be associated with wrong data. Also, accepting poor-quality data into the test database could be responsible for many of the matching errors in biometric systems and could be the greatest weakness of some implementations [4].

Test administrators are essential to the collection of biometric data and metadata that are free of errors. Without data validation and error mitigation, poor-quality data may be allowed into the biometric system. The creation of a test harness is an integral part in the reduction of the error that is introduced by the test administrator. This study focuses on the creation of a test harness and the impact that it has on test administrator error over a two-visit biometric data collection

II. Role of the Test Administrator

Test administrators are an integral part of most data collection systems and commonly make the final decision regarding whether a sample is accepted or not. As such, they need to be trained to understand fully how to handle a system error or problem if it occurs [5].

The test administrator's role is to collect the biometric data, however many times the administrator also ensures that the data collection is performed properly by the subject even if the resulting data are of poor quality. The test administrators' roles vary depending on the test protocol, institutional requirements, and the requirements of the test. For example, in a study by Theofanos et al. [6], test administrators were able to assist subjects to overcome deficits of both video- and poster-based instructional material. In other studies, the test administrator changed the environment [7], tilted and operated the camera [8], ensured that the session proceeded correctly [9], and conveyed complex instructions, while administering the test [10].

Test administrator error could affect many of the data collection procedures, including misuse of a device, inconsistent sampling, providing incorrect instructions to subjects, and incorrect data entry. Some of these issues can be caused by a lack of training, incompetence, overwork, or unrealistic throughput expectations set by the experimenter [4]. Unrealistic throughput expectations could cause test subjects to queue up, which may create additional stress or burden on the test administrator. Hicklin and Khanna [4] recommended the creation of test administrator performance metrics to identify lapses in training and data collection errors. These performance metrics are necessary to fully understand the test administrator's impact on the biometric system.

The reliability of the test administrator is also affected by the length of the test, the test administrator's abilities, the subjects' abilities, and the test conditions where the data collection occurs [11].

One way to mitigate the amount of test administrator error in a data collection environment is with training. Test administrators should be given a set of training topics to introduce them to the biometric technology that is used in the test. Some of these topics include an overview of device operations, how to install the devices, the skills needed to successfully use the devices, start-up procedures, normal operating procedures, human interface procedures, shutdown procedures, and device error response activities [12]. Furthermore, training should take place with an experienced test administrator, and the session should be documented in a quality control manual, such as ISO 17025 [13]. The goal of training is to prevent poor quality from the source [4]. It is important that the test administrator be trained on the actual system that is used during data collection. This training will help to provide experience with the system.

A good test administrator needs to possess many different qualities to function well with complex technological systems and human subjects simultaneously. Complex systems include computerized record systems [9] as well as biometric data [4]. At the beginning of the data collection, the test administrator needs to serve as a host for the subjects. Some of these responsibilities include making the subject feel welcome and making the experience pleasant [14]. These responsibilities can be dependent on the personality of the test administrator. If a study requires many interactions between the test administrator and the subject, an extroverted test administrator may be better than an introverted one. Some studies do not allow the test administrators to talk to the subject, and in these cases, an introverted test administrator may be a better choice.

It is the job of the test administrator to know every aspect of the data collection process and convey any and all necessary instructions to the subjects. An example of this process includes the test administrator giving correct instructions to the subjects for every visit. Dumas and Loring [14] recommended that test administrator duties include greeting subjects, making eye contact, smiling, being relaxed, listening attentively, speaking slowly, and adapting to interaction style. On the other hand, test administrators are advised against acting distracted, using a monotone voice, exhibiting nervousness, rushing subjects, showing annoyance, touching the subjects, and using extraneous technology such as email, or a cell phone, or going on social media.

The workload for a test administrator needs to be balanced so that there is enough work to do without causing the administrator to be overwhelmed by the amount. The test administrator's workload should be monitored so that it does not become a source of quality problems [4]. One report [12] suggests that test administrators should verify test crew demographics and the device installations, conduct system audits, provide the biometric device during conduct of the test, review documentation of daily activities, ensure compliance with test procedures, and validate all collected data.

If the workload is too high, a certain level of automation could be added. Automation aims to provide a system with more capabilities during complex scenarios to take error out of the hands of the test administrator [5]. Automation should primarily be used to eliminate unwanted workload steps such as human data entry, which may prove to be error prone. Unwanted workload includes mental calculations, estimations, comparisons, and unnecessary thinking [15]. Automating these steps will simplify the process for the test administrators, allowing them to focus their cognitive resources on the critical tasks at hand.

It has been reported that fatigue, stress, and distraction are key factors that impact human test administrator performance. With these factors in mind, a person's ability to maintain vigilance and attention reduces over time [5]. Systems should be designed to anticipate test administrator fatigue. Biometric data collection can be a repetitive process, and fatigue will play a role in data collection. Shift workers are even more susceptible to fatigue than are task-based workers. Test administrators are commonly scheduled as shift workers, so fatigue needs to be avoided when possible.

There may also be a link between error frequency and test administrator demand that increases subject waiting times [16]. Additional errors and quality problems can increase with test administrator workload and stress [4]. Many factors will impact the stress levels of a test administrator and may result in an increase in errors or slower throughput times. Ruthruff [17] reported that subjects under a time deadline tend to make more errors in difficult conditions than in easy experimental conditions. The same effect may be observed for test administrators dealing with complex information.

Test administrators are commonly put in situations where scheduling creates time constraints. The need to process subjects through the data collection is crucial, and there is usually a specific time frame in which to do so. This time frame is determined by throughput and budget. If this time constraint passes, additional subjects may start to line up, causing the test administrator to work at an even faster pace so that subjects are not delayed. If multiple subjects come in at the same time, a queue may cause an increase in stress and distraction. The high level of stress and distraction may have a negative impact on decision making by the test administrator.

A common challenge is designing systems that provide functionality but are also easy to learn and use [9]. Some factors that affect the usability of the system include the ergonomic design of the work area, the work station, the Graphical User Interface (GUI), and the user manual. With the test scenario, a GUI should be easy to use and created from the test administrator's perspective. If designed properly, the GUI will help to create a system that is free of confusion. A well-made test administrator-GUI will allow administrators to spend less time searching and thinking, and more time collecting data.

Another principle is to include only the information needed by the test administrator at a given time [15]. Extraneous information should be excluded so that the administrator can focus on the subject and the data collection. Complex systems used in biometric data collections rely on a certain level of test administrator proficiency. Test administrators need to know how to handle the system in the event of a failure.

Qualitative evaluations are used to highlight common errors that occur when the test administrator interacts with a system [5]. These evaluations measure test administrator error rates to improve the system design. Surveys can also be used to learn what problems the system might have from the test administrator's viewpoint. This will help to identify deficiencies in the data collection. Focus groups are used to discover test administrator viewpoints. Focus groups are best used to obtain answers to open-ended questions and acquire as much information as possible from a descriptive answer [5].

Test administrators will affect the data collection procedures. Policy and administration are two key elements of systems management. By implementing best practice policies early on, the biometric system can be designed with cognitive engineering principles in mind. These principles refer to a system that is designed to support the human that is using it [18]. An experiment by Murata and Iwase [15] showed that reaction time was shorter when using cognitively engineered interfaces than when using an interface that was not created using cognitive engineering principles.

III. METHODOLOGY

Prior to developing the test harness, a survey was conducted with current and past test administrators to understand their concerns about biometric data collection. The survey was issued to every test administrator who had worked in a data collection at the testing center between the summer of 2012 and the summer of 2013. The purpose of the survey was to determine the concerns and problems with past tests. The survey contained multiple Likert questions for use as quantitative data. These questions involved degrees of satisfaction with devices, studies, and administrative conditions. The survey also included open response questions for test administrators to write opinions and suggestions.

This survey and a review of literature were evaluated to create the test harness, with the intent of reducing the amount of administrator error. After creation, the test harness was evaluated by conducting pre- and post-data collection on a number of different sensors. The approach for this research was as follows:

- Conduct a survey of test administrators (both current and past).
- Hold a training session based on the ISO 17025 internal quality document between two visits of a biometric data collection to reinforce test protocol.
- Develop the test harness and continuous improvement plan.
- Hold a focus group for test administrators to analyze the test harness.
- Use the test harness on the second visit of the data collection. The second visit occurs approximately nine months after the first visit.
- Analyze the test administrator error results between visits one and two.
- Conduct a post-mortem session, allowing test administrators to provide further recommendations.

IV. RESULTS

A. Test Administrator Survey

Seven test administrators were surveyed about their previous experiences when data collecting. Six of the administrators reported that they preferred consistent schedules. The administrators also reported that they had many different roles during the data collections. In one particular collection, they reported serving as data collector, participant scheduler, test administrator scheduler, error reporter, data manager, and system designer.

Test administrators were also commonly confused on who was in charge of the data collection. In one data collection, test administrators responded by naming three different people who they thought were in charge. A central person in charge will help to create a chain-of-command and provide accountability for procedure changes. Having a single person in charge will also be of benefit for training. For each data collection, the majority of test administrators reported that they received instructions from other test administrators, rather than a central person. This can create fragmented instructions and different behaviors for each test administrator.

Test administrators provided feedback on device specific issues, such as the collection of face data. Test administrators reported that they created errors in past data collections using cameras. Some errors reported included holding the device too close to the subject, using incorrect device settings, and forgetting to perform essential steps. Finally, all seven test administrators reported that they occasionally questioned their own judgments and sometimes forgot at which step of a data collection they were on. This shows the need for standard operating procedures (SOPs) and formal training.

B. *Training of test administrators*

Test administrators were required to read the internal quality manual prior to the second data collection. This manual outlined general requirements to ensure that the administrators complied with internal policies and procedures such as error reporting and best practices. The administrators were trained on how to properly complete all data collection activities. This training session was held prior to the start of visit two and focused on instructing the administrators how they should interact with each station. Before starting the training, each administrator completed a consent form, allowing video to be recorded and their actions to be logged in the database. Test administrators were shown one data collection station at a time and allowed time to practice or ask questions before proceeding. After completing the training, test administrators were shadowed for a minimum of two data collections to ensure that all processes were correctly understood.

Test administrators were also required to pass a quiz demonstrating their knowledge of error reporting procedures and the internal quality manual. A minimum score of 80% (16 of 20 questions correct) was required to work as a test administrator in this data collection. This minimum score was chosen because 80% is typically the score needed for a "B" grade. After completing the quiz, each test administrator was given her or his score, along with the justification for each incorrect answer, in private. The most commonly missed question involved when improvements could be implemented to the data collection. Four of the test administrators thought changes could not be made during the data collection because changes could jeopardize the results. Test administrators were reminded of the importance of continuous improvement and that changes could be made during the data collection as long as they did not affect the integrity of the data. Individual test administrator results are shown in Table 4.1.

TABLE 1. TRAINING QUIZ RESULTS

Test Administrator	% Correct
1	90%
2	80%
3	95%
4	80%
5	80%
6	95%
7	100%

C. *Improvements to the test harness*

Following the feedback from the test administrator survey, the review of literature, and the data from the first visit of the study, the following improvements were made to the test harness:

1. Test administrator shift log-in provided additional data for the project manager on logged hours to monitor for fatigue, and test administrator accountability for errors. Prior to starting the data collection work shift, the test administrator logged into the database with unique credentials created during training. If a previous test administrator had not logged out of the database, a "Switch User" button allowed the administrator to change to her or his account. The primary function of the database GUI was to remind the administrator of common operations that must be completed at the start of the day and before each new subject.

2. Housekeeping checklists – checkboxes were introduced for common items at the start of the day and at the start of each subject

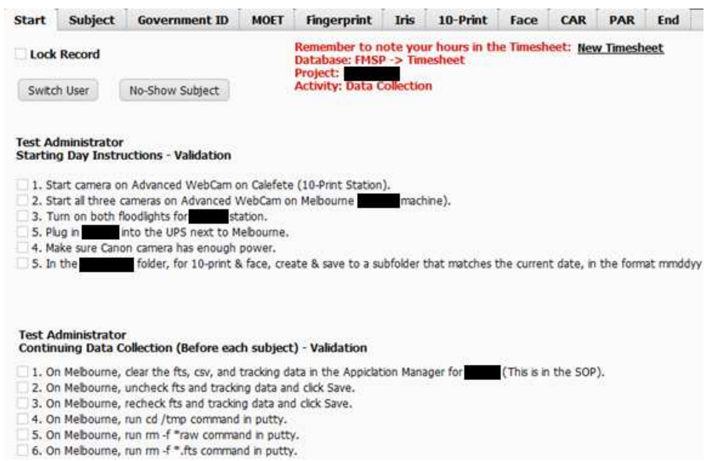

Fig 1: Test Administrator Checklists

Checkboxes were used to ensure that operations such as clearing system log files, turning on additional lighting for the iris recognition station, and starting video recordings were all completed. Items noted in the checkboxes were operations that were commonly forgotten during visit one. The checkboxes were not always used to validate that the housekeeping activates were performed by test administrators. Although the first set of checkboxes was only performed at the start of the day, test administrators needed to validate that all six of the activities listed under "Test Administrator Continuing Data Collection (Before each subject) – Validation" were completed before each new subject arrived. Test administrators did not use these checkboxes for 10 of the 81 subjects, but it is likely that these activities were still completed.

After verifying that the operations were completed, the test administrator navigated to the “Subject” tab to check-in the subject. In the event of a subject missing an appointment, the “No-Show Subject” button was selected to create a report of this event in the database. This report justified that the subject should not receive compensation for the day and that data were not lost or accidentally deleted by the test administrator.

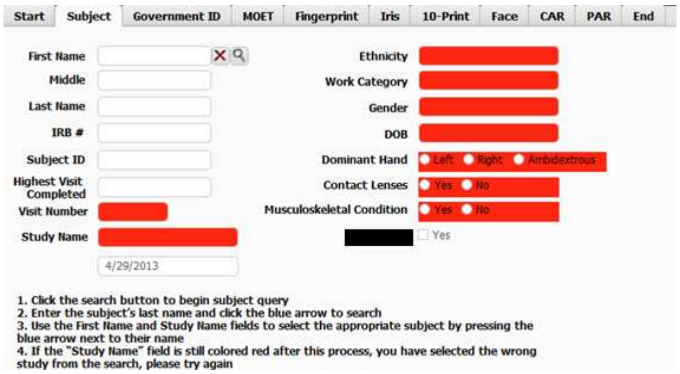

Fig 2. Subject Tab

3. Validation of metadata. Test administrators were not able to move to the next screen if there were outstanding data requirements

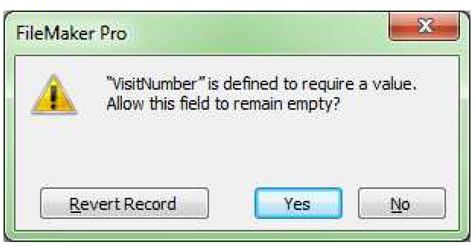

Fig 3. Empty Field Error

A new screen in the test harness prompted test administrators to look up a past subject or add a new subject to the database. As these subjects were returning for their second visit, any demographic information that missing or incorrectly collected from the first visit was highlighted in red to prompt the test administrator to collect this information from the subject. These fields, such as a standardized format for data of birth, were automatically validated by the database.

4. Sensor-specific data collection improvements. To aid in the complexity of a 10-stage data collection, the test harness provided guidance to the test administrators. This guidance included all necessary hardware and software instructions, scripts to be read to the subjects, and a place to document any deviations from the test protocol. In the first visit of the study, 12 subjects were missing their fingerprint metadata. Visit two, required all subjects to have this data collected and store, and also improved the capture of sebum (oiliness of a fingerprint), as it was collected incorrectly during the first visit. A change to the test protocol showed a significant reduction in error. Table 2 shows the improvements of the fingerprint metadata between visits.

TABLE II. FINGERPRINT METADATA BETWEEN VISITS

Metric	**Visit One**	**Visit Two**
Missing Subjects (All Fields Blank)	12	0
Temperature (Blank)	0	0
Skin Texture (Blank)	0	0
Pigmentation (Blank)	0	0
Sebum (Measured Incorrectly)	99	0
Moisture (Blank)	0	0
Elasticity (Blank)	0	0
Skin Color (Blank)	0	0
Keratin (Blank)	0	0
Total Erroneous Fields	195	0

Another area that was improved was the data collection of face data. Improvements were gleaned from the test administrator survey. An additional template was used to align the subjects’ faces, and this improved the image quality of the face metrics as measured by a commercially available face quality tool. Samples in all but one category increased or stayed consistent at 100%. The height to width ratio underwent a substantial improvement from 50.93% to 100%. The camera template also aided to improve the degree of blur by helping the test administrators to line up the subjects’ faces. The full results of the face improvements are shown in Table 3.

5. Introduction of preventive action and corrective action reports. To create a philosophy of continuous improvement, Corrective Action Request (CAR) forms were implemented in the test harness. These database forms were completed by test administrators to recommend process changes after an error had occurred. Preventive Action Request (PAR) forms were also implemented. PARs were completed by the test administrators to recommend a process change before an error occurred. These forms included a unique identification number, the source of the problem, the urgency of the problem, a description of the problem, proposed actions to fix the problem, who assigned the

problem, the assignee, and whether the problem had been corrected or not. PARs and CARs were built into a tab in the software so they could be easily accessed by test administrators. Upon submission, the CAR or PAR was stored as a database record and a copy was emailed to the test administrator to whom it was assigned. It became the responsibility of the assignee to complete the request or forward it to another test administrator who could.

TABLE III. FACE QUALITY METRICS BETWEEN VISITS

	% Compliant	
Metric	**Visit One**	**Visit Two**
Eye Separation	95.34%	97.21%
Eye Axis Angle	97.21%	99.20%
Eye Axis Location Ratio	87.58%	97.61%
Centerline Location Ratio	0%	0%
Height to Width Ratio	50.93%	100%
Head Height to Image Height Ratio	97.52%	97.61%
Image Width to Head Width Ratio	69.26%	37.85%
Eye Contrast	100%	100%
Brightness Score	100%	100%
Facial Dynamic Range	100%	100%
Percent Facial Brightness	100%	100%
Percent Facial Saturation	100%	100%
Degree of Blur	60.56%	68.13%
Image Format	100%	100%

D. *Post-mortem*

A post-mortem session was held three weeks after the conclusion of the data collection. Six test administrators were asked about their experiences and opinions, on the benefits, scope, schedule, costs, quality, communication, staffing, risks, and action requests of the data collection. Test administrators unanimously reported that the database implementation and additional tools made data collection easier than past collections. Although there were scripts provided, some of the test administrators thought that they were repetitive, and therefore did not adhere strictly to the script. It was suggested that scripts should vary per visit, or possibly contain only a list of guidelines that should be addressed. Test administrators reported that there were still difficulties in recording all of the comments from the subject, especially when the data collection was complex. Test administrators suggested that a second test administrator would make this process easier and more accurate.

From the perspective of the new test harness, it was clear that the improvements were beneficial. All test administrators indicated that they found the harness to be straightforward and easy to use, although there were still some difficulties with the CARs and PARs. Some test administrators reported that it was not always clear to whom the request should be assigned.

It was recommended that future studies should continue to use the electronic data collection procedures and they should continue to be located in the software suite for quick referencing. If the study were to be repeated, test administrators unanimously agreed that hardware and systems should be upgraded to fully optimize the established test harness.

V. CONCLUSIONS

This study examined the creation of a test harness to measure and reduce the amount of test administrator error in a biometric system. By introducing new training methods and creating database functionality to help test administrators, the amount of erroneous fields in the data collection was reduced. A continuous improvement philosophy was also built into the framework to aid future studies as well. The post-mortem activity was used to help in future studies by addressing remaining issues. In this framework, the use of CARs and PARs was essential to improving data collection and should continue to be used to provide accountability and also to keep records for the funding agency. Over the course of the data collection, a total of 44 corrective action requests and five preventive action requests were filed and resolved by the test administrators.

This study has shown that the test administrator plays an important role in the integrity of a subject’s biometric data. Without a system to log test administrator actions, it is difficult to determine whether an error was caused by the subject, the test administrator, or an extraneous factor. The amount of erroneous data fields and non-compliant face samples was greatly reduced by paying specific attention to the role of the test administrator. Standardized training and error reporting were key in instructing test administrators how to correctly collect data, as well as how to solve any issues that may occur.

The test harness used in this data collection is fully modifiable. The methodology of the test administrator data collection suite can be adapted for any future biometric data collection. Data entry fields can be easily altered based on the nature of future tests. The ability to look up a subject in the database will also continue to help establish a database of subject data and metadata to aid in future research.

REFERENCES

[1] International Organization for Standardization, “ISO/IEC 2382-37:2012 - Information processing systems - Vocabulary - Part 37: Harmonized Biometric Vocabulary,” 2011.

[2] J. Campbell and M. Madden, “ILO Seafarers’ Identity Documents Biometric Interoperability Test (ISBIT-4) Report,” 2009.

[3] T. C. Redman, “Poor Data Quality on the Typical Enterprise,” *Communications of the ACM,* vol. 41, no. 2, pp. 79–82, 1998.

[4] A. Hicklin and R. Khanna, “The Role of Data Quality in Biometric Systems,” *White Pap. Mitretek Syst. (February 2006)*, no. February, pp. 1–77, 2006.

[5] I. Graves, M. Butavicius, V. Macleod, R. Heyer, K. Parsons, N. Kuester, A. Mccormac, P. Jacques, and R. Johnson, "The Role of the Human Operator in Image-Based Airport Security Technologies," *Def. Sci. Technol. Organ.*, vol. 338, pp. 147–181, 2011.

[6] M. Theofanos, B. Stanton, S. Orandi, R. Micheals, and N. Zhang, "Usability Testing of Ten-Print Fingerprint Capture." NIST, Gaithersburg, p. 56, 2007.

[7] E. P. Kukula, S. J. Elliott, R. Waupotitsch, and B. Pesenti, "Effects of illumination changes on the performance of geometrix face vision 3D FRS," in *38th Annual 2004 International Carnahan Conference on Security Technology, 2004.*, 2004, pp. 331–337.

[8] M. Theofanos and B. Stanton, "Usability Testing of Face Image Capture for US Ports of Entry," *Biometrics Theory, …*, 2008.

[9] A. W. Kushniruk, V. L. Patel, and J. J. Cimino, "Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces.," *Proc. AMIA Annu. Fall Symp.*, pp. 218–22, Jan. 1997.

[10] International Biometric Group, "Comparative Biometric Testing Official Test Plan," 2002.

[11] H. K. Sarmah and B. B. Hazarika, "Determination of Reliability and Validity measures of a questionnaire," *Indian J. Educ. Inf. Manag.*, vol. 5, no. 11, pp. 508–517, 2012.

[12] Transportation Security Administration, "Plan for Biometric Qualified Products List (QPL)," 2005.

[13] International Organization for Standardization (ISO)., "ISO/IEC 17025:2005 General requirements for the competence of testing and calibration laboratories," 2005.

[14] J. Dumas and B. Loring, "Moderating Usability Tests." Elsevier, pp. 1–24, 2008.

[15] A. Murata and H. Iwase, "Effectiveness of Cognitively Engineered Human Interface Design," vol. 20, no. 5, pp. 7–10, 1998.

[16] A. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: A review of applications, methods and models," *Eur. J. Oper. Res.*, vol. 153, no. 1, pp. 3–27, Feb. 2004.

[17] E. Ruthruff, "A test of the deadline model for speed-accuracy tradeoffs," *Percept. Psychophys.*, vol. 58, no. 1, pp. 56–64, 1996.

[18] D. Norman, *User Centered System Design: New Perspectives on Human-computer Interaction*. Hillsdale, NJ: CRC Press, 1986, p. 544.