National Center for Border Security and Immigration

Research Lead: The University of Arizona (Tucson, Arizona)

Field Tests of an AVATAR Interviewing System for Trusted Traveler Applicants

Department of Homeland Security – Grant No. 2008-ST-061-BS0002

"This research was supported by the United States Department of Homeland Security through the National Center for Border Security and Immigration under grant number 2008-ST-061-BS0002. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security."

THE UNIVERSITY OF ARIZONA

Report Authors

Jay F. Nunamaker – University of Arizona

Elyse Golob – University of Arizona

Douglas C. Derrick – University of Nebraska at Omaha

Aaron C. Elkins – University of Arizona & Imperial College, London

Nathan W. Twyman – University of Arizona

jnunamaker@cmi.arizona.edu, egolob@cmi.arizona.edu, dderrick@unomaha.edu, aelkins@cmi.arizona.edu, ntwyman@cmi.arizona.edu

Executive Summary

Overview

The National Center for Border Security and Immigration (BORDERS) conducted two field tests of an automated interviewing system for trusted traveler interviews. Customs and Border Protection's (CBP) trusted traveler programs expedite and facilitate the arrival of pre-approved, low-risk individuals into the United States. The process for obtaining a trusted traveler permit is not easily scalable because it includes an in-person interview, a requirement that is time-consuming and human resource intensive, resulting in long wait times for applicants. Automating trusted traveler interviews can help decrease personnel demands while simultaneously providing a supervisory officer with additional information on the applicant's potential risk.

AVATAR Automated Interviewing

BORDERS has developed the Automated Virtual Agent for Truth-Assessment in Real-Time (AVATAR), a kiosk-like system designed to automatically and independently conduct natural credibility assessment interviews. AVATAR uses a virtual conversational agent to conduct interviews while simultaneously detecting potential anomalous behavior via analysis of data streams from noninvasive sensors such as cameras, microphones, and eye tracking systems. Potential indicators of deception are compared to an individual baseline - individuals are not flagged for simply being nervous about the interview. To ensure privacy, all data has been kept anonymous and only aggregate data is reported.

AVATAR Field Trials

Prior to the herein reported field studies, these automated screening systems had not been tested in a real-world environment or process. The field tests were conducted at the SENTRI Enrollment Center in Nogales, Arizona. The adapted system for conducting trusted traveler screening interviews successfully conducted more than 200 interviews.

Summary of Results

The research aims of the field tests ranged from specific requirements gathering to examining overall operational feasibility:

- Operational Feasibility
- Applicant Acceptance
- Officer Acceptance
- Speech Recognition
- Dialogue Processing
- Language
- Question Wording and Flow
- Decision Support Interface
- Interface Height
- Document Processing

The field tests provided many valuable insights that will inform future AVATAR designs, including needs for improved speech recognition, important improvements for the user experience, user interface design modifications, and many other operational insights. See Table 1 below.

Table 1. Summary of Results and Future Directions

Research Aim	SENTRI Field Tests Results	Next steps
Operational feasibility	▪ Operational feasibility validated ▪ 258 interviews completed ▪ High rate of user acceptance	▪ Exposure to wider range of applicants ▪ More officer feedback ▪ Extended vetting of operational fit
Applicant Acceptance	▪ AVATAR well received by users in both field tests	▪ Exposure to a wider range of subjects and emotional responses ▪ Evaluate user acceptance in alternative contexts via surveys and interviews
Officer Acceptance	▪ Formal dress of the AVATAR in pilot 2 improved perceptions	▪ Additional requirements gathering for alternative enrollment centers ▪ Formal officer feedback via anonymous survey and collaboration sessions
Speech Recognition	▪ Background noise and users who spoke too quickly caused problems ▪ Speech recognition software improved ▪ Hardware and software adjusted to improve performance	▪ Further improvements to the speech recognition and noise filtering algorithms ▪ Exposure to a variety of accents and operating conditions (e.g. ambient noise) will aid in improving the speech recognition system
Dialogue Processing	▪ Additional instructions added to improve interaction ▪ Minor delay between questions truncated user responses (fixed in field test 2)	▪ More robust dialogue ▪ Additional dialogue management system modifications underway
Language	▪ During pilot 1, AVATAR spoke only English ▪ Spanish added as second language for pilot 2 (selected by 47% of applicants) ▪ Level of Spanish adjusted to match regional usage	▪ Improvements to Spanish dialect ▪ Integration of new languages and dialects ▪ Automated selection of language
Question Wording and Flow	▪ Some questions in test 1 were reworded based on officer input ▪ Observations of real world interactions provided valuable insight into how officers communicate with applicants	▪ Officers will be interviewed to determine not only the wording, but also the goal of questions ▪ Terms used by officers to clarify the questions need to be integrated into the AVATAR script ▪ Increase dialog flow flexibility
Decision Support Interface	▪ Initial tablet interface was difficult to navigate ▪ Tablet interface redesigned to meet officer needs	▪ Further user interface improvements useful across domains ▪ More detailed explanations of risk assessment scores will be integrated into the interface
Interface Height	▪ Some applicants were too tall or too short for the system to be used effectively	▪ AVATAR kiosks currently under development will automatically adjust to accommodate applicants of all heights

Document Processing	▪ Self-service document scanning and automatic document analysis was not tested in the SENTRI field tests ▪ Observations suggested that the AVATAR might be used to automate document collection using a self-service scanner	▪ Investigate how documents might be best captured and processed by the AVATAR ▪ Explore how AVATAR can perform document validity checks automatically

Future Directions

Lessons learned from these field trials will be instrumental in moving the AVATAR system from the proof-of-concept stage closer to a viable government and industry tool for automated screening and risk assessment. In order to improve operational performance and examine risk assessment performance in an operational environment, additional field tests will be necessary. Whereas the reported field tests focused on an initial assessment of operational feasibility, future field tests will assess the performance of the credibility risk assessment algorithms, with particular emphasis on determining the optimal number and type of sensors, the appropriate risk segmentation rates, and the most optimal questioning protocol. Quantification of process improvements from using multiple kiosks and/or automating biometrics collection or document scanning will be an important area of investigation, as will be the optimization of the user interface for both applicants and decision makers. From a technical standpoint, speech recognition and noise filtering technologies performed adequately in quiet, enclosed spaces, but are as yet untested in more open settings. Eye tracking and kinesic recognition technologies also need testing in noisier field environments.

Background

CBP's trusted traveler programs expedite and facilitate the arrival of pre-approved, low risk individuals into the United States. The program allows speedy processing of low-risk individuals, providing benefits for travelers as well as for CBP. Membership makes it faster and easier for millions of visitors and business people to enter the country at Ports of Entry (POE) and to have more predictability while traveling. For CBP, officers can focus their limited resources on unknown, higher-risk travelers seeking to enter the country. Trusted traveler programs include Global Entry at select U.S. international airports, SENTRI at the U.S.-Mexico border, and NEXUS at the U.S.-Canada border. Dedicated lanes allow members to pass quickly through border checks.

Trusted traveler programs are based on the concept of risk segmentation. Applicants submit background information through the Global Online Enrollment System (GOES). CBP uses this information to conduct thorough background checks via criminal, customs, immigration, agricultural, and terrorist databases. Those who pass this background check complete a personal interview with a CBP officer to verify that all information is correct, and to conduct a final assessment of the individual's risk. During this interview, biometric checks are also used to identify travelers and to check again in the above-listed databases. If the applicant passes this interview, he or she receives a Radio-Frequency Identification (RFID) document that is used to authorize more rapid travel through border points of entry.The appeal of expedited travel has proven to be strong. In recent years, applicant demand has increased exponentially, rapidly outpacing staffing and infrastructure levels. This pattern has resulted in long delays in scheduling interviews. As a result of these backlogs, CBP has been unable to keep up with demand using the current system.

AVATAR as a Trusted Traveler Interviewing System

BORDER has developed Automated Virtual Agent for Truth-Assessment in Real-Time (AVATAR), a kiosk-like system designed to automatically and independently conduct credibility assessment interviews [2]. These autonomic screening systems use virtual conversational agents to conduct interviews while detecting potential anomalous behavior via analysis of data streams from noninvasive sensors such as cameras, microphones, and eye tracking systems [1].

As noted above, the major bottleneck in the trusted traveler application process is the in-person interview to verify information and complete the enrollment process. Through two methods, integration of AVATAR systems can decrease human resource demands while increasing the ability to assess potential risk. First, the kiosks can decrease the human resource requirement through automating part of the processing, especially the interview. Automation using multiple kiosks concurrently can have a force multiplier effect, automatically processing low risk individuals while referring those judged a relatively higher risk to an officer for questioning. Second, the kiosk risk segmentation output can serve as a decision support, helping officers focus follow-up questions on topics that the AVATAR identified as triggering anomalous responses from the applicant. Through these two methods, automated screening kiosks can potentially increase efficiency and performance simultaneously.

This paper reports on two field tests where an AVATAR system was employed as part of a trusted traveler application process. Prior these field tests, the AVATAR had not been tested in a real-world environment or process. BORDERS actively sought to enhance understanding of operational and theoretical nuances through the application in a working environment. In pursuit of these, BORDERS conducted several briefings to DHS stakeholders on this research. In response to one of these outreach activities, David P. Higgerson, Director of Tucson Office of Field Operations (OFO) requested a follow-up meeting with his staff. In summer 2010, BORDERS researchers met with Director Higgerson, several port directors, and other specialists. OFO suggested that a small test site for AVATAR at a POE would be ideal for BORDERS for the purpose of gaining a more detailed understanding of how the AVATAR could assist with this need. Several follow-up meetings served to identify an ideal field test application in the trusted traveler program at the Nogales Enrollment Center located at the DeConcini POE. We worked extensively with OFO in Nogales, Arizona to create an automated interview for the Secure Electronic Network for Travelers Rapid Inspection (SENTRI) program, integrate it within the current process, and evaluate its performance.

Due to CBP time constraints, only six weeks were available between the identified site and the initial field test. Also, limitations on interviewing space and project funding prevented testing more than a single AVATAR. The infrastructure, funding, and short window for preparation limited the scope of the field tests to an operational feasibility and requirements gathering exercise. Process efficiency improvements (e.g., force multiplier effects) resulting from using multiple kiosks and performance evaluation of risk assessment flags were reserved for follow-up studies.

The pilot studies were funded by DHS Science and Technology Directorate Office of University Programs as part of on-going research activities. As such, the proposed tasks were outlined in a supplemental project (Appendix B). The following five tasks were specified and successfully completed:

Task 1: Customization of the AVATAR Kiosk for SENTRI Interview Questions

Task 2: Configuration of Kiosk to Deliver Results to CBP Officers

Task 3: Transportation, Delivery, and Installation of the AVATAR Kiosk Device in Nogales, AZ

Task 4: Pilot Field Test for SENTRI Enrollment Center Operations

Task 5: Data Analysis

Methodology

Two field tests were conducted to evaluate AVATAR in a realistic setting. The focus of the field trials was on operational feasibility and gathering knowledge not easily obtainable through interviews alone. The novelty of the system concept is such that requirements can be difficult to identify without hands-on use. Once the concept and field setting were identified, the process of requirements gathering, system development, evaluation, and refinement progressed iteratively. The first field test relied on a more limited set of system features and its purpose was to identify key areas for improvement to be incorporated in the second field test. The reported final results thus reflect both operational performance and the most important lessons learned as a result of the two field tests.

Requirements Gathering

A series of site visits to the SENTRI Enrollment Center at the DeConcini POE served to provide a high-level understanding of the SENTRI program and process. Requirements were also gathered via visits with key stakeholders in Washington, D.C., such as John Wagner (CBP Executive Director of Admissibility and Passenger Programs), and Colleen Manaher (CBP Office of Field Operations Executive Director of Planning, Program Analysis, and Evaluation), and visits to the Tucson Office of Field Operations. These visits occurred prior to, during, and after the field tests. As the project progressed, site visits became more pointed, detailed, and specific to particular processes. For the sake of brevity, the full SENTRI process will not be documented here. A high-level view of the process and where the automated screening system was inserted is depicted in Figure 1.

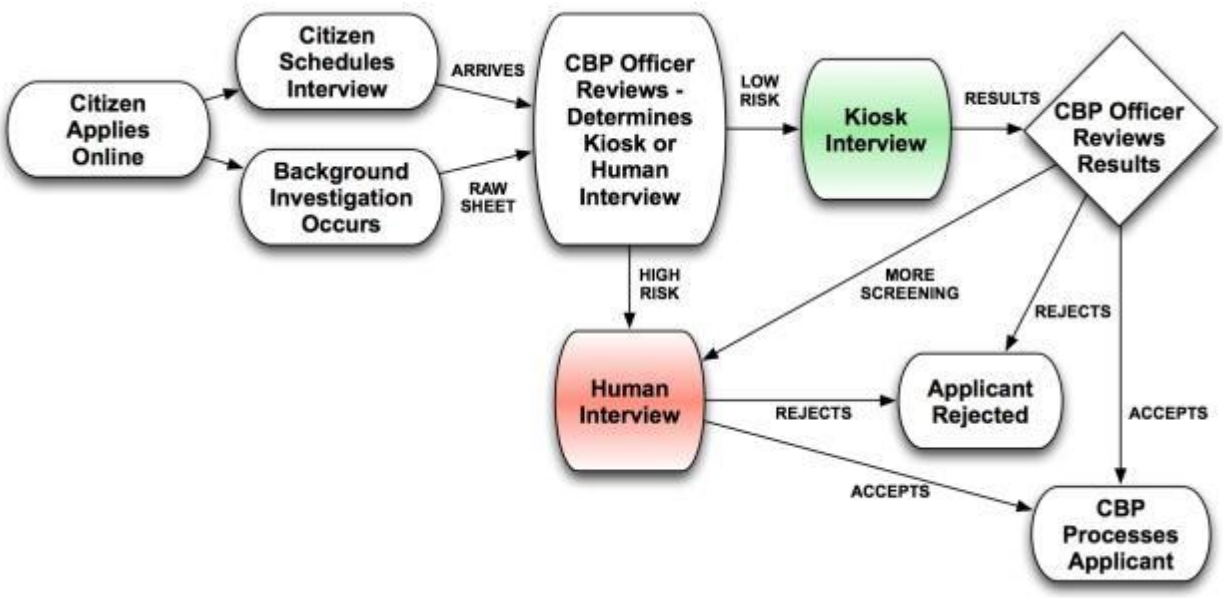


Figure 1. Process Flow for Field Tests

Location

Both field tests were conducted at the SENTRI Enrollment Center in the DeConcini POE in Nogales, Arizona where the AVATAR operated in a real-world, real stakes environment. CBP officers occupy several offices adjacent to a medium-sized waiting area for applicants. Each officer conducting SENTRI interviews is located in a private office with access to CBP databases as well as a fingerprint scanner.

AVATAR interviews were ultimately conducted in a second-floor office. We determined that the optimal positioning for the AVATAR would be in one of the private offices used to conduct standard SENTRI interviews. Placing the AVATAR in a common area or near the waiting room would likely introduce data collection challenges for vocalic sensors due to a high level of ambient noise. Furthermore, we determined that requiring applicants to interact with the kiosk in an open, public area might heighten their arousal/stress as they could be observed by others, and could introduce privacy concerns if answers to sensitive questions were overheard.

Interview

An officer asks 20 standard questions of each applicant, many of which are yes/no questions. These questions are based on the responses provided in the on-line application form. Based on preliminary observations and staffing of the AVATAR, we estimated that SENTRI interviews last an average duration of 20 minutes, with each officer conducting roughly 8-10 interviews a day (depending on staffing and scheduling). Each interview involves document review, questions, fingerprints, and a photograph. If there are discrepancies, the same officer can ask follow-up questions. However, asking too many follow-up questions of one applicant can increase the wait time for other applicants, and/or decrease the amount of time available for questioning other applicants. The rejection rate for interviewees is low, since the information from the online application has been verified prior to the interview.

To ensure applicant privacy during the field tests, the officer interface only presented data in aggregate format (e.g., explained that vocal features suggested uncertainty, rather than displaying raw vocal data). BORDERS assumed full control and responsibility for field test data; no data from the field tests were recorded or stored by CBP.

Officer Interface

Since the 20-question interview process is routine, it lends itself well to automation. In consultation with CBP, we determined that the AVATAR should ask the 20 questions, record responses, and display them to an attending officer. Questions that elicited behaviorally anomalous responses would be flagged for the officer who could then probe more deeply into these topics.

Broad variations in personality and mood were expected to be present among the applicant population. Thus, anomalous responses needed to be identified only based on an individual baseline—no individual measurements would be compared to a population norm. In this way, simple nervousness or stress about being in an interview would not cause an individual response to be flagged.

Automating the SENTRI Interview

Four key steps were required to adapt the AVATAR from an earlier experimental prototype [2] toward a trusted traveler interviewing context: incorporating speech recognition, creating a risk decision algorithm, generating SENTRI-specific questions, and developing a user interface.

Incorporating Speech Recognition

First, speech recognition was incorporated to ensure a more seamless interaction. Previous iterations required a button press after each response, but in evaluations pre-dating the field tests, this proved confusing for many applicants. Speech recognition without training (i.e., without a phase gathering specific baseline audio) is a daunting problem. To simplify the issue, the kiosk was purposefully limited to searching for three words reliably (“yes”, “no”, and “repeat”) with no training. Other spoken words were ignored or designed to trigger an “I didn’t understand” response.

As in similar iterations [2], a virtual agent (i.e., a talking avatar) controlled by intelligent agent software conducted screening interviews. Speech recognition allowed the agent to conduct a dynamic interview, taking one of several branches in the script depending the answer given.

Implementation of a Risk Decision Algorithm

Based on our previous research [1], many potential sensor inputs could have been employed as decision criteria. Ideally, an interaction is designed to take advantage of several behavioral and physiological indicators that an individual is expressing uncertainty, is hiding something, or has increased cognitive or emotional arousal. The indicators are calculated based on the subject’s physiological and behavioral reactions to questioning (not necessarily what they say) and are captured using a variety of instruments. For these field tests, we focused exclusively on vocalic features. This decision stemmed from a perceived rigidity of the number and type of questions the AVATAR could ask, and the shortened preparation period prior to Field Test 1. Both factors limited the scope of what could be incorporated in the risk assessment algorithm. Vocalic features were ultimately chosen above other possibilities because of relative ease of integration and prior success using vocalic features to assess credibility. Vocalic features were recorded and used to flag responses according to their relative risk level.

Anomaly detection was not based on population norms but a personal baseline. This means an individual would not be considered more risky if he or she simply found the experience stressful or was apprehensive about the results. Rather, very large deviations from their own personal "normal" were flagged as anomalies.

Figure 2. AVATAR Kiosk used in the SENTRI Field Trials

Generating SENTRI-specific Interview Questions

The standard questions asked in SENTRI background interviews were adapted for the kiosk. Open-ended questions were revised to one or more yes/no questions. Following the first field test, question wording was revised based on officer and interviewee feedback and observer notes. To ensure clarity of speech, expert vocalists were employed to record AVATAR voices. Because of time constraints, the questions were only available in English for the first field test, but Spanish versions of these and similar questions were developed and vetted during and after the first field test. Selections from the final interview script follows:

Opening:

> "Hello, I am AVATAR, and I am conducting a trial program with the University of Arizona for trusted traveler applicants. I am now going to ask you certain questions regarding your SENTRI application. In a moment, I will ask you to say your name, and then answer some yes or no questions. Please wait for me to finish speaking before saying 'yes' or 'no.'"

1. Have you ever used any other names?
2. Were any of the other names used for illegal purposes?

3. Do you live at the address you listed on your application?
4. Have you lived at the same address for the last five years?
5. On your application, did you list all addresses at which you have lived within the last five years?
6. Did you accurately list all your employment activities?
7. Have you visited any foreign countries in the past five years?
8. On your application, did you list all foreign countries you have visited in the last five years?
9. Was any of your foreign travel for a purpose other than business, vacation, shopping, or visiting family or friends?
10. Have you ever used illegal drugs?
11. Do you normally travel through the Port of Nogales when entering the U.S.?
12. Have you ever violated any U.S. customs, immigration, or agricultural laws?
13. Do you understand the trusted traveler program requirements?
14. Do you understand that any violation of program requirements will be dealt with more severely because of the "trusted traveler" status that been placed on you?

Final: "Thank you. Please wait for further instructions."

Following the first field test, researchers conducted a site visit to the Nogales POE to collaboratively review the performance of the question script with officers. Prior to field test 2, the question list was revised for understandability and the question flow was adjusted. Feedback from the officers and managers effected many changes in both question wording, questions asked, and question flow. A graphical representation of the question flow of the automated interview is shown in Figure 3. The numbers in the nodes of the graph do not correspond to the questions listed above.

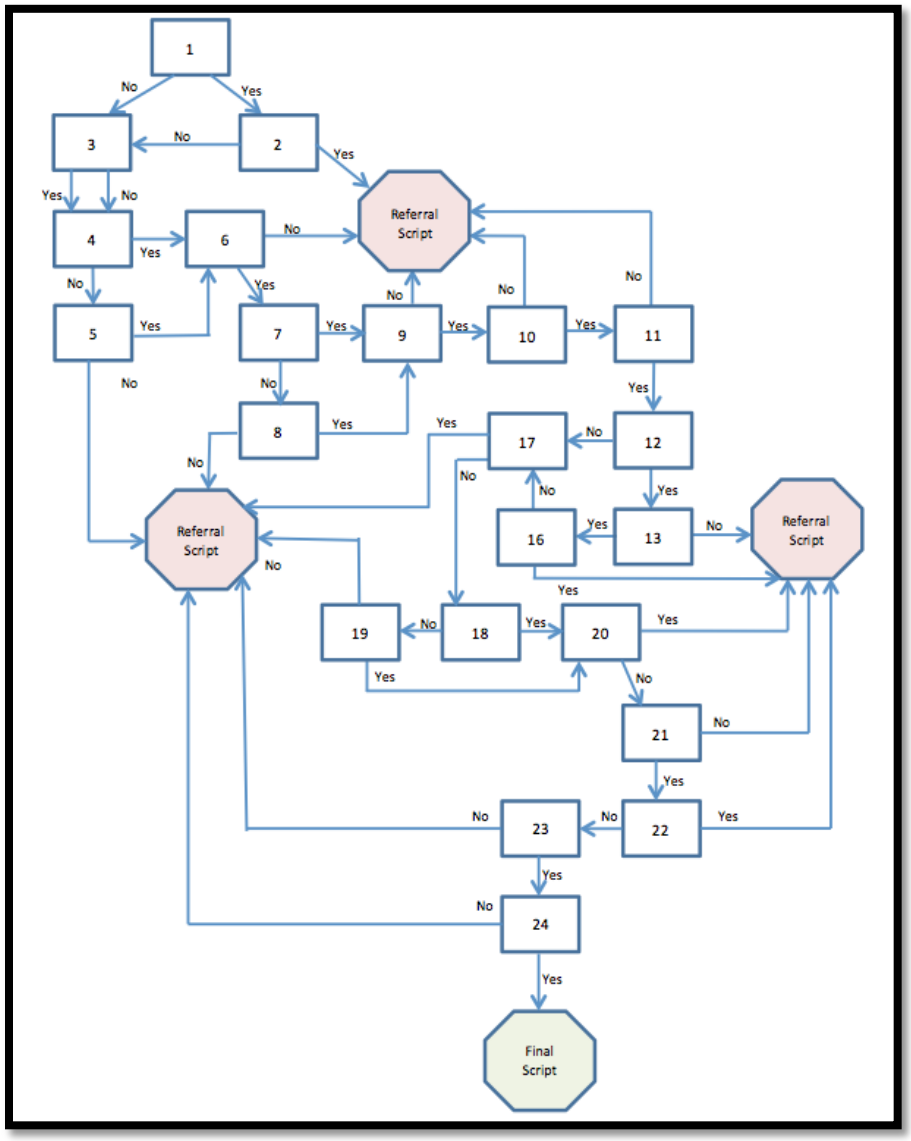


Figure 3. Dynamic Flow of the Automated SENTRI Interview Questions

Development of the Officer Interface

We also worked with CBP to understand how the resulting interview information could be best presented to officers. Integrating results with their current systems was not feasible for a field trial, thus a separate interface viewable via a tablet was developed to display the results. This application effectively provided officers with answers to each question and a risk assessment score (via color coding responses) for each response indicating potential topics for follow-up questioning. If the AVATAR detected anomalous readings in one of these factors, the response is coded orange. If more than one variable is flagged, the response is coded red. Normal readings are coded green. As noted previously, all data was standardized on individual baselines, so anomalies were not based on overall stress level or nervousness. A screenshot of the main screen shown to a CBP officer following an interview is displayed as Figure 4.

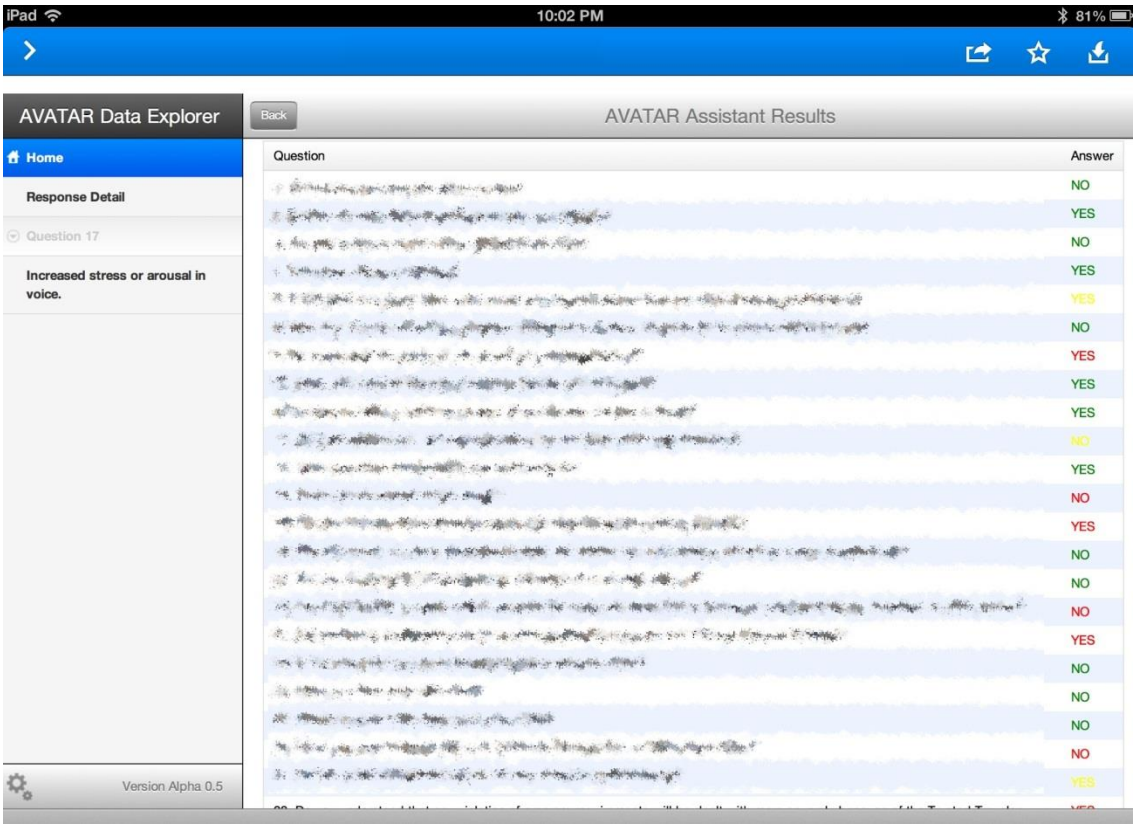


Figure 4. Screenshot of iPad Application for SENTRI Automation. Question text redacted.

Field Test 1

The AVATAR interview was conducted as the first phase of the standard SENTRI interview. Applicants were informed that the AVATAR was a research initiative conducted by BORDERS at the University of Arizona, and that participation in the field test was voluntary. If willing to participate, applicants then began the interview with the AVATAR. At the end of each interview, officers were able to view responses and risk assessment flags using the tablet interface, which provided insight as to whether the officer should follow up on a given question during the standard interview. As this was a field test and not part of the official enrollment process, the officer overseeing the interview ultimately made the decision as to which questions would be followed up on. In no way did the AVATAR determine whether an applicant's SENTRI application was accepted or denied.

As stated in the background section, time and resource constraints limited the ability to examine the performance of the risk assessment algorithms. The automation and proper functioning of the AVATAR system within the SENTRI process was the focus of this project.

Figure 5. Avatar Kiosk Deployed for Field Trial in Nogales, AZ

Performance Results

Prior to the first field test, 37 pilot interviews were conducted on site using research staff to ensure proper calibration and system operation. During field test 1, the AVATAR conducted a total of 134 interviews with actual SENTRI applicants. Of those interviewed, 93 told the AVATAR they were American citizens, 22 Mexican citizens, and 5 indicated they were neither. Fourteen were flagged before being questioned about their citizenship status. Question responses and interview progression (i.e., which questions were asked and how many were interviews successfully completed) were captured to determine success rate. Vocalic data were successfully captured and used to generate a risk assessment score in all completed interviews.

Anomaly Detection Results

66% (N=1398) of all questions answered received a risk rating of green (Low), 31% rated orange (Medium High), and 3% were rated red (high risk). Only one question was rated as yellow (medium risk). There was no statistical difference in the proportion of risk levels assigned to American or Mexican citizens, $\chi^2(6, N = 2{,}2094) = 3.31$, $p = .77$. This indicates no bias in risk assignment by citizenship.

Field Test 1: Lessons Learned

The key goal of this project was to observe the AVATAR in an operational context, thereby gathering requirements and identifying key operational nuances and limitations that could not be discovered through traditional requirements gathering interviews with operational and strategic personnel. Several such lessons were learned from the first field test. First, a high rate of successful interviews confirmed the operational feasibility of the AVATAR-based SENTRI interview. Second, no SENTRI applicants refused to participate, and none expressed hesitation toward talking to a machine.

The system performed adequately, though limitations will need to be addressed prior to a full test and evaluation of a near production-level prototype. Beyond the issues created by time and resource constraints (detailed in the background section), key limitations included speech recognition, process, and user interface limitations.

Most of the field study 1 limitations were addressed by improvements made to the prototype which were implemented between field studies.

Speech Recognition Limitations

Most of the speech recognition problems stemmed from interviewees speaking too quickly (i.e., talking over the AVATAR before the question was completed and the microphone activated). As a result, the initial instructions were revised to request that the interviewee wait until a question was finished before answering. The transitions between question and answer were thus sharpened, allowing for faster onset of voice recording after each question.

In the first round, the AVATAR only communicated in English, which was a concern for some applicants who spoke limited English. Spanish speaking capability was added to the next prototype, along with a selection screen for the individuals to choose their preferred language.

Process Limitations

Another lesson from the field trial concerned the interview script (see Figure 2). Several of the questions were deemed confusing by participants, and were reworded based on officer input. Initially, the AVATAR was programmed to end the interview when a question was answered in a disqualifying manner (e.g., if an individual admitted to entering false information on the application). This was in accordance with current CBP processes, but it did not allow collection of the full data from an interaction. The abrupt exit of the interview was deemed suboptimal because in many cases, more information would have been helpful for the officer. In coordination with CBP, the next iteration of the system conducted the full interview regardless of disqualifying answers. This allowed for a full interview, better data collection, and more useful information for CBP officers.

Some stakeholders suggested the AVATAR could use a more professional appearance. For field test 2, we had the avatar wear a shirt and tie.

Officer Interface Limitations

The tablet interface was somewhat difficult to navigate, and available information wasn't as easily accessible as was desired. The officer interface was therefore revised to reduce the number of clicks needed to access relevant information.

Field Test 2

A second field test was conducted to examine the effectiveness of the system revisions, and further evaluate feasibility. The second field test was conducted in the same location and manner as the first. As in the first field test, participation by applicants was strictly voluntary, and AVATAR output was not used by officers' decision processes. A large number of interviews were desired and projected, however physical space constraints at the CBP facility necessitated a reduced workload than the AVATAR was capable of processing. Because the AVATAR had to be located in a working office rather than a space all its own, only one officer could use the AVATAR on any given day. Scheduling conflicts with available space also limited the number of days available at the Nogales Enrollment Center for field test 2. The AVATAR system thus had to be removed due to space constraints.

Performance Results

A total of 124 SENTRI applicants underwent questioning by the AVATAR in field test 2. Of those interviewed, 79 were American citizens, 38 Mexican citizens, and 7 indicated they were neither. One of the improvements made was inclusion of a Spanish speaking AVATAR: 46.7% of the applicants chose to complete the interview in Spanish.

Table 2. Interview Counts for Field Tests 1 and 2

	American Citizens	Mexican Citizens	Neither	Unknown	Total
Field Test 1 Interviews Conducted	93	22	5	14	134
Field Test 2 Interviews Conducted	79	38	7	0	124
Total	172	60	12	14	258

Anomaly Detection Results

The previous field test revealed that four risk levels were too many, from the officer's perspective. We reduced the risk levels to only three for field test 2. The proportion of questions rated as green (low risk) was 61% (N=960), 35% (N=553) were rated yellow (medium risk), and 4% were rated red (high risk). Replicating the first field trial, the vocal risk scores were not proportionally different among applicants with different citizenships (Mexican, American, Other), $\chi^2(4, N = 1{,}577) = 3.25$, $p = .52$. There also was no statistical difference on risk assignment for applicants that completed the interview in either English or Spanish, $\chi^2(2, N = 1{,}577) = 1.38$, $p = .50$. Based on the only demographic factors made available, there was no culture bias in the risk assignment during the interview.

Field Test 2: Lessons Learned

Several additional lessons were learned from the second field test, including the need for a height-adjustable design, further voice recognition improvements, additional process automation, and formalization of a performance feedback structure.

Height-Adjustable Design

The eye tracking system equipped with the kiosk was not used in the field study, but has potential application for tracking gaze patterns and pupil dilation during certain strategic questions. However, to activate this sensor would require each individual be a similar height, or have a height-adjustable interface.

Voice Recognition and Noise Filtering

Though improved, there were several cases where speech recognition did not work. Further refinements of the untrained speech recognition algorithms will be necessary going forward. For instance, dynamic gain adjustment was an identified feature needed for avoiding signal clipping. Speech recognition and audio processing is of particular interest because the AVATAR was in an enclosed room with little noise. Future tests involving multiple co-located kiosks may experience serious difficulty in this area. Ultimately either major system improvements will need to be made, or each kiosk may need to be in a separate enclosed space.

Process Automation

The AVATAR system successfully automated the standard questioning portion of the interview. However, other portions of the application process including instruction and document scanning could also be automated. These functions also would have potential to both conserve human resources and increase risk assessment score reliability.

Performance Feedback

When observers were present, any errors in the process, whether technical or procedural were easily identified and documented. In future field tests, a formal feedback reporting process for success or failure of each interview will be a necessary testing process improvement. This will allow more nuanced identification of system weaknesses.

Conclusions

The AVATAR system for Trusted Traveler background screening interviews successfully underwent field trials. Lessons learned from these trials will be instrumental in moving the AVATAR system from the proof-of-concept stage closer to a viable commercial tool for automated screening and risk assessment. Based on the lessons learned from the field tests, there are several design enhancements planned for future iterations of the AVATAR, as well as additional research necessary prior to a full test and evaluation phase. The key areas of investigation for future field tests are summarized in Table 1.

In addition to improvements for the CBP Trusted Traveler program, other potential uses for the AVATAR include 1) TSA Pre-Check, 2) CBP new hires and periodic reinvestigations, 3) USCIS applications (citizenship, asylum, refugee, etc.), and 4) Department of State visa adjudications.

Acknowledgements

We would like to acknowledge the cooperation of the CBP officers at DeConcini Port of Entry/SENTRI Enrollment Center in Nogales, Arizona, as well as those at the Tucson Office of Field Operations. Funding for this research was provided by the US Department of Homeland Security through the National Center for Border Security and Immigration (BORDERS; grant number 2008-ST-061-BS0002). However, any opinions, findings, and conclusions or recommendations herein are those of the authors and do not necessarily reflect views of the US Department of Homeland Security. The views, opinions, and/or findings in this report are those of the authors and should not be construed as an official US Government position, policy, or decision.

References

[1] Nunamaker, J.F., Jr., Burgoon, J.K., Twyman, N.W., Proudfoot, J.G., Schuetzler, R., and Giboney, J.S., "Establishing a Foundation for Automated Human Credibility Screening", in 2012 IEEE International Conference on Intelligence and Security Informatics (ISI), 11-14 June 2012, pp. 202-211.

[2] Nunamaker, J.F., Derrick, D.C., Elkins, A.C., Burgoon, J.K., and Patton, M.W., "Embodied Conversational Agent–Based Kiosk for Automated Interviewing", Journal of Management Information Systems, 28(1), 2011, pp. 17-48.

Appendix A: Summarized Pilot Study Field Notes

This section contains field notes from observers of the SENTRI AVATAR pilot tests. The notes are not comprehensive, but are summarized to eliminate replication and to group like content.

Phase 1 Field Notes - December 2011 - January 2012

Question Content

- The question and branching structure had been defined before deploying the AVATAR. However, several issues were found.
- The question about school surprised people. Not everybody was asked about this when filling out a question, so people were unsure what to answer. This question should have only been asked to students.
- For employment, it is possible that it is not applicable. Perhaps a branching structure could be added to account for this.
- Some questions refer to "SENTRI," but people might be enrolling for "GLOBAL." This could cause confusion.
- The foreign country question confused some people. Foreign countries should be understood as any country besides the USA and Mexico. For example, during an interview, one Mexican citizen stopped and asked if the USA was a foreign country.
- The questions were difficult for some people to understand, especially when English is a second language.

Speech Recognition

- There were many instances where people gave a "yes" response but the AVATAR heard "no" and vice versa.
- Several people tried to answer the question before the AVATAR was listening. Some people repeated their answer after a few moments of silence, but others had to be prompted to repeat their answers.
- Sometimes the background noise was registered as a "yes" or "no" response.
- Sometimes, responses such as "That's a good question" would be interpreted as "yes" or "no" responses.
- Some people would respond, "No, sir."
- Some people tried to explain why they answered "yes" or "no."
- Rebooting the AVATAR corrected at least one case where the audio was not being recorded.

Kiosk Environment and Management

- There was a significant amount of background noise in the room where the interviews took place, but the AVATAR performed well.
- One woman stood to the side and quietly helped her husband (who mostly spoke English, but not perfectly). Those showed up as slow responses, but her talking didn't interfere with the system understanding his commands.
- Agents did not consistently put the AVATAR in sleep mode overnight.

- Sometimes there would be multiple people in the screening room. One person would be taking the interview while another person reviewed SENTRI guidelines with the officer. However, when an AVATAR interview was occurring, everybody felt that they needed to be silent.
- The microphone had to be recalibrated several times.

User Interface

- The red button often has to be pushed twice for the interview begins. The first push brings up the AVATAR's face. The second push starts the interview. Several people pushed the button once, then waited for further instruction.
- One woman realized that she should have answered a question differently, but had no to way to go back and correct her response.
- When the AVATAR said, "Hello," one woman responded, "Hi." This did not throw of the interview, but perhaps indicated an expectation of a deeper dialog.
- A Spanish speaking male did the interview while somebody translated for him. He answered "Si" several times instead of "Yes."
- For the iPad, it was suggested that the interviews be listed by date in descending order so that the latest interview was at the top.
- One man hit the red button after answering a question because he interpreted the delay incorrectly.
- Many applicants were unable to use the AVATAR because they only spoke Spanish, or because they were children.

Officer Considerations

- The same officer observed most AVATAR interviews. When another office filled in for her, it was clear that the new office was unfamiliar with the AVATAR, its purpose, or how to use it correctly.
- The agents did not consistently use the iPad to see which questions were flagged by the AVATAR. Some officers did not know that there was an ipad application to go with the kiosk.
- Officers have to do a lot of paperwork which they feel interferes with their ability to detect deception.
- Officers often told the applicants how to respond.
- It appeared that some officers viewed the AVATAR more as an academic experiment rather than a tool that could assist them with their jobs.

Interviewee Comments

- One gentleman wanted the AVATAR to be more stern.
- One woman seemed concerned that the AVATAR was a lie detector.
- One gentleman said it was "weird" to interact with a 3D model.
- A woman said that the system was "cool."
- A Spanish speaking woman said it was easy. She said that it didn't make her nervous; she was more curious than anything because she hadn't done it before.
- One woman called the AVATAR "creepy."

- One woman thought that the AVATAR's eyes were "weird" and "distracting.'

Phase 2 Field Notes - August 2012

Questions

- Some were confused about the question about whether they always entered the USA through Nogales. One woman came in through Mariposa frequently, which according to the officer observing, falls under the umbrella of the Nogales Port of Entry, so the woman should have answered, "yes.'
- Questions about using different names confused people who changed their last names because of marriage.

Speech Recognition

- Generally, the speech recognition seemed to improve from Phase 1.
- Several people had to repeat answers.
- People who speak softer than normal are not always understood.
- One woman was asked to repeat things nearly a dozen times. The AVATAR interview was ended early because of the difficulty.
- The AVATAR seemed to get stuck on the same question and ask it over and over.
- Background noise did not seem to affect the AVATAR.
- The speech recognition seemed to work better when people spoke normally instead of trying to over-emphasize words.

Kiosk Environment and Management

- Currently, the kiosk sits in a single officer's office. Other officers seem reluctant to bring their interviewees into the other agent's office.

User Interface

- The Spanish interview allowed many more people to use the AVATAR than in Phase 1.
- Many responses are being flagged as suspicious.
- For example, in one interview, 16 of 21 responses were marked as suspicious.
- The iPad requires the officer to click on each response that is flagged, which may require an officer to click on a single interview a dozen times.
- The first interview of the day sometimes show up with the previous day's date.
- If possible, it would be good to dynamically adjust the microphone sensitivity to match the interviewee's speaking volume.

Officer Considerations

- The officers do not seem to know how they are supposed to use the AVATAR. They have not been formally trained.
- One officer did not know that there were different reasons why a response might be flagged.

- After a few days, some of the officers had to be reminded how to turn on the iPad.
- Officers are unfamiliar with the terms reported on the iPad. For example, to one officer, "voice quality" was just a measure of how well the AVATAR's microphone heard the response.
- The officers' perception of what cues indicate deception do not always match what the AVATAR looks for. For example, officers trained to look for certain behavioral cues might not focus on vocalics when conducting an interview. When they see the AVATAR's analysis based on vocalics, they may ignore that information.
- Officers asked people to perform the interview with the AVATAR *after* having conducting the interview with the applicant face to face. In these cases, the AVATAR is duplicating work instead of reducing it.
- Many officers saw the AVATAR on the news, and that seemed to pique their interest. Some days had lots of officers stop by to see the AVATAR in person.
- One officer commented that the AVATAR's eyes are "harsh" and that his forehead is too big.
- Though the AVATAR was supposed to be used by all officers doing interviews, it seemed that only the officer who had the kiosk in her office really used it.

Interviewee Comments

- One older gentleman was skeptical of the AVATAR.
- One woman commented that she had just read about the AVATAR.

Appendix B: Original Proposal for the AVATAR Field Tests

Supplemental Project: RA1-1.3a

AVATAR Kiosk Pilot Test[1]

Jay F. Nunamaker, Elyse Golob – University of Arizona

Douglas C. Derrick– University of Nebraska at Omaha

jnunamaker@cmi.arizona.edu, egolob@cmi.arizona.edu, dderrick@unomaha.edu

Project Abstract

In YR 4, BORDERS will begin a 6- 8 week field trial of its AVATAR kiosk technology for deception detection in Nogales, Arizona. This pilot field test, scheduled to begin on December 9, 2011, will consist of a limited trial of AVATAR kiosks as screening entities in a border crossing scenario. In addition, it will serve as a proof-of-concept solution for a current issue that Customs and Border Protection (CBP) faces. The Secure Electronic Network for Travelers Rapid Inspection (SENTRI) program provides expedited processing for pre-approved, low-risk travelers wishing to enter the U.S. In order to qualify for the program, applicants must complete an on-line application, voluntarily undergo a thorough biographical background check against criminal, law enforcement, customs, immigration, and terrorist databases including a 10-fingerprint law enforcement check and a personal interview with a CBP officer. This personal interview is a time-consuming activity for CBP officers, and as the program volume increases, it will require more and more officers to staff the enrollment centers. Given current budget constraints and the demand on officers' time, an AVATAR kiosk can serve as a force multiplier that will free up officers' time to conduct higher-level tasks. These kiosks can ask the standard interview questions to applicants and provide real-time feedback to CBP officers. The information provided to the officers is based on the AVATAR kiosk sensor data, which measure cues of deception that are not perceptible by human senses. Using this technology, one officer may be able to monitor 4-8 AVATAR kiosk stations at a time.

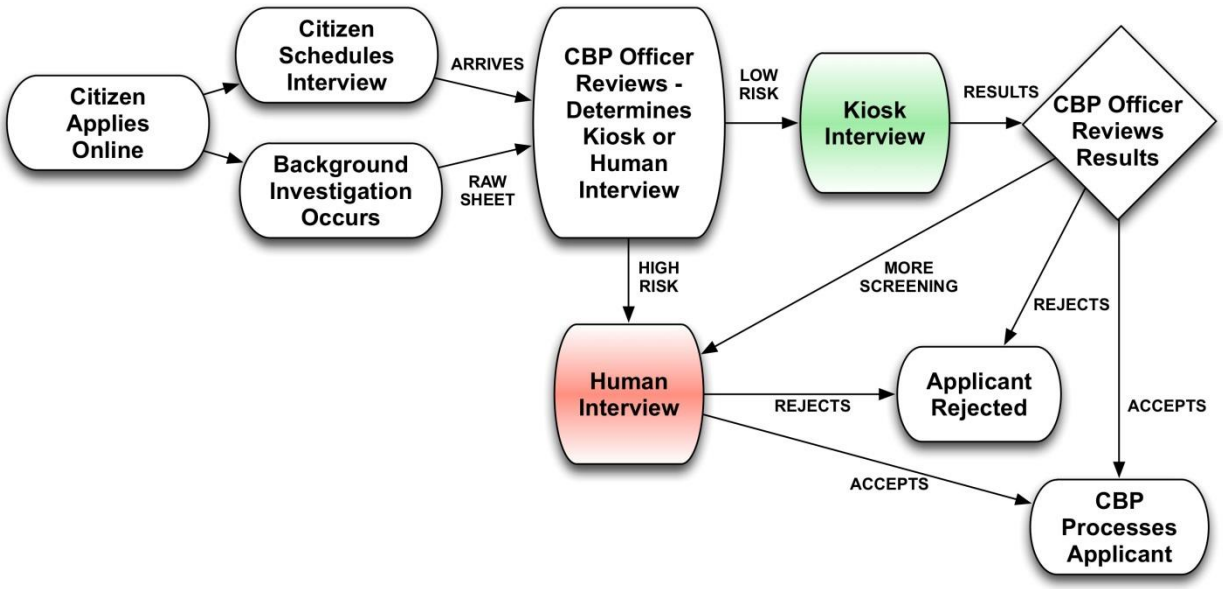
[1] "This research was supported by the United States Department of Homeland Security through the National Center for Border Security and Immigration (BORDERS) under grant number 2008-ST-061-BS0002. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security."

Significance to DHS

Automated agents (AVATARs) have the potential to greatly assist DHS by freeing personnel to focus on mission-critical tasks. This field test aims to help improve the effectiveness of the SENTRI enrollment process because AVATAR kiosks can be replicated and function as force multipliers to alleviate the traffic load on human officers. Furthermore, as programs such as SENTRI and other trusted traveler initiatives expand, the current model of one officer conducting one interview at a time is unsustainable. The AVATAR does not get fatigued and can perform the standard interviews multiple times with the same level of vigilance. Finally, the AVATAR can detect cues of deception and malicious intent that are not perceptible to human senses. The object of this field study is to expand the capacity of one officer so that he or she can oversee multiple applicants at the same time, while ensuring accuracy and improving human decision-making.

Research Description

This field test will test the AVATAR's capabilities as an interviewer for the SENTRI program. In order to enroll in the SENTRI program, applicants fill out an on-line application and schedule an interview at an Enrollment Center (EC) to verify the information. The EC is an indoor, controlled environment located near a Port of Entry. An officer asks 20 standard questions of each applicant, most of which are yes/no questions. Each interview generally takes 20 minutes for questions, fingerprints, and photograph. An officer then collects the application fee. If there are discrepancies between the online application and interview, the officer can ask follow-up questions. However, this increases the wait time for other applicants. The rejection rate for interviews is low, since the information from the on-line application has been verified. We propose to insert the AVATAR kiosk into the process as shown below:



From an operations management perspective, the current bottleneck in SENTRI application processing is the CBP personal interview. Using the AVATAR kiosk will alleviate this bottleneck and keep human decision makers in the decision loop. The first proof-of-value field trial will only use one kiosk. However, it is expected that one officer will be able to process the output of multiple kiosks, and the feasibility and value of the AVATAR kiosks will be explored in this initial field trial.

The research will focus on providing an innovative way of reducing the mundane workload of officers by conducting the standard 20-question interview. Since the interview process is routine, it easily lends itself to automation and frees up officers for higher-level functions, including follow-up questioning for inconsistencies. The field test will be conducted at the SENTRI EC in Nogales, AZ where the AVATAR will operate in a real-world, real stakes environment. The research project consists of 5 overall tasks.

Task 1: Customization of the AVATAR Kiosk for SENTRI Interview Questions

For this task, we will tailor the AVATAR kiosk interaction and incorporate the SENTRI interview questions. The task involves rendering the questions with the AVATAR, testing the system, monitoring the sensor output and updating the software code for this particular application. It will also include implementing interview branches based on the real-time recognition of affirmative or negative responses.

Task 2: Configuration of Kiosk to Deliver Results to CBP Officers

The AVATAR kiosk must be able to deliver the results of the interaction in real-time to the CBP officer. The task includes incorporating a secure web server that will allow the delivery of the results to the computer used by the CBP officer.

Task 3: Transportation, Delivery, and Installation of the AVATAR Kiosk Device in Nogales, AZ

We will deliver the kiosk to the Enrollment Center and install the device in coordination with CBP. Initial system testing will occur to make sure that the device is ready for the field test.

Task 4: Pilot Field Test for SENTRI Enrollment Center Operations

We will conduct a several week long field trials as outlined in the research description.

Task 5: Data Analysis

Based on the outcomes of the field trials, we will document the results of the AVATAR kiosk system including officer and applicant perceptions, kiosk performance, and recommendations for system improvements. The report will outline the steps ahead for a wider deployment and test of the AVATAR kiosk technology.

Methodology

We use iterative rounds of research-based design and experimentation, leveraging existing technologies and developing new ones in order to create a flexible kiosk framework that works reliably in controlled settings and in the field. Design science serves as a useful framework because of its focus on creating, and then evaluating, information technology solutions intended to solve identified organizational or information problems. We follow the design science research methodology and its seven guidelines, which include: design as an artifact, problem relevance, research rigor, design as a search, and research communication.

Because many components of the AVATAR were developed in a laboratory environment, much of the emphasis for this pilot test will be on design evaluation. Although there are several methods of design evaluation, this project will utilize observational and analytical methods. For the observations, a case study will be performed in which the AVATAR is studied in- depth in the operational environment. Because the AVATAR is an innovative product in early production form, descriptive methods of analysis are appropriate for much of the evaluation. However, dynamic analytical analysis will also be performed to determine capabilities such as performance and accuracy.

To help us evaluate the design of the AVATAR, we will observe real applicants in the field as they use the AVATAR. These observations should be extremely valuable in pinpointing issues that should be addressed in future iterations. The observations will be recorded as field notes by members of the research team. The field notes will contain any and all data relevant to the AVATAR, including system accuracy, user praise and complaints, and any challenges that arise. These observations will demonstrate the value of real-world field testing as part of the design evaluation within the design science methodology. Valuable insights can be gained through field testing that are unavailable in a controlled laboratory setting. The pilot field test of the AVATAR will allow us to better understand both the real-world value and real-world challenges associated with its potential implementation at the border.

Student Involvement

Students will be used in the software development and kiosk implementation. Specifically, students will help with Spanish language translation, AVATAR voice implementation, hardware installation and configuration, and software updates. Students will also be used to facilitate the field trial.

Transition Strategy

The field trial is directly related to the transition efforts of the AVATAR kiosk and will allow us to make design refinements for a wider deployment of the technology. We expect that this trial will last for several weeks in an operational environment with real customers and DHS personnel. We have already identified manufacturers of the kiosk, the infrared sensor, and the stereoscopic camera. We will explore the possibility of partnering with a business to integrate the field ready kiosks and to manage the fielding of the kiosks for an expanded test.

The basic transition plan is shown below:

- Initial field trial in operational environment (next 3 months)
- Design refinements (6 months after initial trial)
- Commercial partner identified (next 6 months)
- Creation of multiple field ready kiosks (next 12 months)
- Initial field test in limited locations (completed in next 18 months)
- Design refinements (21 months)
- Wider dissemination of kiosks, pending results of field test (24 months)

All of the steps outlined above will be coordinated with DHS and advancement to each new phase is dependent on results of the previous phase, funding availability and DHS approval as appropriate.

Milestones

Task	Milestones
1	Customization of AVATAR for SENTRI enrollment questions
1	Configuration of kiosk to allow data delivery to mobile devices
2	Real-time processing of applicant responses
2	Real-time results delivered to the mobile device
3	Transportation, Delivery, and Installation of kiosk to Enrollment Center
4	Field trial of AVATAR Kiosk
5	Data analysis and report