Submitted 5/15; published X/XX

Evaluating Trustworthiness of AI-Enabled Decision Support Systems: Validation of the Multisource AI Scorecard Table (MAST)

Pouria Salehi PSALEHI@ASU.EDU
Yang Ba YANGBA@ASU.EDU
Nayoung Kim NKIM48@ASU.EDU
Ahmadreza Mosallanezhad AMOSALLA@ASU.EDU
Anna Pan AROLSO10@ASU.EDU
Myke C. Cohen MYKE.COHEN@ASU.EDU
Yixuan Wang YWAN1290@ASU.EDU
Jieqiong Zhao JZHAO153@ASU.EDU
Shawaiz Bhatti SABHATT1@ASU.EDU
Arizona State University, USA

James Sung JAMES.SUNG@HQ.DHS.GOV
DHS Office of Intelligence and Analysis, USA

Erik Blasch ERIK.BLASCH.1@US.AF.MIL
Air Force Office of Scientific Research, USA

Michelle V. Mancenido MVMANCENIDO@ASU.EDU
Erin K. Chiou ECHIOU@ASU.EDU
Arizona State University, USA

Abstract

The Multisource AI Scorecard Table (MAST) is a checklist tool based on the U.S. Intelligence Community's analytic tradecraft standards to inform the design and evaluation of trustworthy AI systems. In this study, we investigate whether MAST can be used to differentiate between high and low trustworthy AI-enabled decision support systems (AI-DSSs). Evaluating trust in AI-DSSs poses challenges to researchers and practitioners. These challenges include identifying the components, capabilities, and potential of these systems, many of which are based on the complex deep learning algorithms that drive DSS performance and preclude complete manual inspection. Using MAST, we developed two interactive AI-DSS testbeds. One emulated an identity verification task in security screening, and another emulated a text summarization system to aid in an investigative task. Each testbed had one version designed to reach low MAST ratings, and another designed to reach high MAST ratings. We hypothesized that MAST ratings would be positively related to the trust ratings of these systems. A total of 177 subject matter experts were recruited to interact with and evaluate these systems. Results generally show higher MAST ratings for the high-MAST conditions compared to the low-MAST groups, and that measures of trust perception are highly correlated with the MAST ratings. We conclude that MAST can be a useful tool for designing and evaluating systems that will engender trust perceptions, including for AI-DSS that may be used to support visual screening or text summarization tasks. However, higher MAST ratings may not translate to higher joint performance, and the connection between MAST and appropriate trust or trustworthiness remains an open question.

1. Introduction

Decision-making is increasingly dependent on artificial intelligence (AI) in many high-stakes domains, such as healthcare, where AI systems are used for tasks like assessing breast lesions or diagnosing arrhythmia (Phillips-Wren, 2012; Zhu, Gilbert, Chetty, & Siddiqui, 2022). These AI-enabled decision support systems (AI-DSSs) help institutions meet high service demands with limited human resources (Knop, Weber, Mueller, & Niehaves, 2022). However, alongside these advancements, there is a growing concern about the trustworthiness of AI-DSSs, particularly in safety-critical areas such as national security and medical diagnostics wherein AI errors could result in catastrophic consequences (Cooke & Durso, 2007). In these contexts, sensitive applications of AI technologies are typically designed with humans-in-the-loop.

Human-in-the-loop systems often integrate human supervision with AI decision-making processes to ensure that decisions are not only data-driven but also contextually informed and ethically sound (Parasuraman & Wickens, 2008). Trust plays a pivotal role in these systems, because trust influences the willingness of individuals to engage with and rely on the AI. People's trust in automation has been found to be closely linked to their confidence in the system's performance and its consistency with human values (Lee & See, 2004). In safety-critical and time-constrained task environments, trust in AI becomes crucial when individuals must rely on AI recommendations or actions without consistent monitoring or the ability to intervene. Consequently, trustworthy AI-DSSs with humans-in-the-loop need to maintain a delicate balance. These systems should enable human supervisors to effectively intervene in situations where the AI may underperform or commit errors, particularly in edge cases, while allowing for the AI's reliable application in routine tasks. Achieving this balance is critical in scenarios where both safety and rapid decision-making are paramount.

Existing design frameworks and best practice guidelines for human-in-the-loop systems often offer broad recommendations (e.g., de Visser, Peeters, Jung, Kohn, Shaw, Pak, & Neerincx, 2020; Schaefer, Chen, Szalma, & Hancock, 2016), presenting a challenge in translating these recommendations into specific, implementable features in AI technologies. This has raised concerns about the practicality and impact of these frameworks in the development, testing, and evaluation processes of AI systems. There is a continuous need for more precise guidance that can be quickly and effectively operationalized to optimize the design-test-evaluation cycles, an ongoing pursuit for both researchers and practitioners.

To bridge some of these existing gaps in the design cycle, the Multisource AI Scorecard Table (MAST) was developed as a structured checklist to aid in designing and evaluating AI systems for trustworthiness (Sung, Nguyen, Blasch, Daniel, G, & Mason, 2019; Blasch, Sung, & Nguyen, 2021). MAST is grounded in the principles of the Intelligence Community Directive (ICD) 203, which sets forth nine tradecraft standards for evaluating the quality of human intelligence reporting in the Intelligence Community (ODNI, 2015). These standards include sourcing, uncertainty, distinguishing, analysis of alternatives, customer relevance, logical argumentation, consistency, accuracy, and visualization. MAST extends these criteria to include aspects of data transformation, aggregation, labeling, data display, and contextual relevance that cover various phases of the AI system's life cycle from data collection to continuous monitoring (Blasch, Sung, Nguyen, Daniel, & Mason, 2019). The underlying premise of MAST is that by integrating these nine criteria into AI system design, the outputs become more transparent and trustworthy, thereby improving the AI's utility and effectiveness in human-in-the-loop systems. While MAST's usefulness has been demonstrated through several case studies in intelligence and reconnaissance tasks (Sung et al., 2019; Blasch et al., 2021), empirical studies dedicated to validating this tool are yet to be published as of this writing.

This paper pursues this objective by applying the MAST framework to the design and subsequent evaluation of two AI-DSSs. Facewise is an identity verification system and READIT (REporting Assistant for Defense and Intelligence Tasks) is a system for text summarization and data visualization. The primary goal of this study is to investigate two key aspects of MAST framework, namely, (1) the potential of the MAST to aid in the design and evaluation of human-AI systems that reflect human trust perceptions, and (2) the broader applicability of MAST in assessing the trustworthiness of AI-DSSs for other safety-critical task environments, extending beyond its use in intelligence or reconnaissance tasks.

Our research offers valuable insights into the utility of MAST as a tool for the design and evaluation of AI systems, while also contributing to the existing body of knowledge on trust in technology. Our findings suggest that integrating the nine MAST criteria into AI system design positively influences users' trust perceptions. Moreover, we find that MAST is effective in improving trust perceptions not only in systems designed for intelligence tasks but also in a broader range of AI-enabled applications. However, the study also uncovers potential limitations of MAST, suggesting areas for future research. An important finding, echoing similar findings in other research, shows that high trust perceptions and in this case high MAST scores also do not necessarily translate to higher human-AI system performance.

This study underscores the challenge of operationalizing universal criteria that can improve human-AI system performance and that can effectively incorporate trust concepts into human-AI system design. Despite these challenges, our findings support the potential of MAST as a viable tool in system design. It contributes to aligning design with practitioner norms, facilitates the documentation of essential transparency information, and can engender high trust perceptions in systems intended for safety-critical tasks.

2. Background

The role of people as the final arbiters over imperfect automation has a long history (Sheridan, 1975; Bainbridge, 1983). In the supervisory control structures that govern most human-AI systems, people are tasked with assessing and, if necessary, intervening in AI outputs. However, many DSSs are designed for task environments in which people rarely have the cognitive and physical resources to sufficiently understand, assess, and intervene with every recommendation (McGuirl & Sarter, 2006). This is especially true in safety-critical systems, in which people may be expected to attend to every outcome produced by imperfect AI-DSSs.

Limitations in human decision-making amid imperfect AI-DSSs have resulted in novel types of problematic outcomes, some of which have been catastrophic. For example, people tend to overly rely on decisions recommended by automation or AI, even when there are clear indications that the recommendation may be wrong (e.g., automation bias; Skitka, Mosier, & Burdick, 1999). An infamous case is from the Iraq war, in which the Patriot missile system's DSS erroneously identified allied fighter jets as enemy aircraft. Operators of the missile system approved the DSS-recommended decision to attack the aircraft, causing the fratricide of American and British pilots (Cummings, 2006). More recently, a series of wrongful arrests in the United States has been traced to law enforcement reliance on facial recognition technologies that have considerable racial and gender biases (e.g., Hill, 2020; Hill & Mac, 2023). However, upon recognizing errors in AI recommendations, there is also a tendency for people to reject future AI recommendations (e.g., automation aversion; Dietvorst, Simmons, & Massey, 2015), especially by experts in the decision-making domain (Snow, 2021).

People's tendency to overuse, misuse, or disuse DSS has long been linked to poorly calibrated perceptions of the DSS's trustworthiness with respect to its actual reliability (Parasuraman & Riley, 1997). As such, methodological frameworks, policy guidelines, and other tools for designing and evaluating DSS trustworthiness have proliferated alongside advancements in AI-DSS capabilities. These include but are not limited to, the Microsoft UX Design Principle (Microsoft, 1995), NISTIR 8330 by National Institute of Standards and Technology (Stanton & Jensen, 2021), AI Fairness 360 Toolkit by IBM (Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, Nagar, Ramamurthy, Richards, Saha, Sattigeri, Singh, Varshney, & Zhang, 2019, and others), IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (Chatila & Havens, 2019), UXPA Guidelines for Trustworthy User Experiences (Kriskovic, Dutta, & Brewer, 2017), or Ethical OS Toolkit (Lilley, Currie, Pyper, & Attwood, 2020). Although these tools do not all explicitly focus on the concept of trust and trustworthiness, they share an underlying motivation that the design, development, and evaluation of AI systems that impact people and organizations require attention to human factors.

Despite the existence of many frameworks and tools to guide the design of trustworthy AI and other software systems, designing for trust and evaluating trustworthiness in practice remains a challenge. There is a wide translation gap between theory and practice, partly because trust is an abstract construct with myriad closely related concepts. For example, designing trustworthy systems also often involves designing for transparency, individual differences, workload, situation awareness, and attending to other possible factors like etiquette and anthropomorphism (Hoff & Bashir, 2015; Parasuraman & Miller, 2004). Another challenge to effectively designing trustworthy AI is that the various expert communities in different domains may define trust differently. These differences in definitions can be attributed to what each community values most and therefore, designing for trustworthy AI means something different for every community. For example, the intelligence community might value high-quality data as a foundation for high-quality analysis. For the transportation security community, it might value high-quality decisions made at the front lines that could affect traveler safety, more so than data integrity.

To address this gap between concept and practice of designing and evaluating the trustworthiness of AI systems, the Multisource AI Scorecard Table (MAST; Sung et al., 2019; Blasch et al., 2021) was developed by the AI Team of the 2019 Public-Private Analytic Exchange Program, supported by the Office of the Director of National Intelligence and Department of Homeland Security. MAST describes nine criteria derived from analytic tradecraft standards ICD 203 to assess the trustworthiness of intelligence reporting, and additionally includes a four-level quantitative breakdown for each criterion. The idea is that MAST could serve as an easy-to-use checklist for designing trustworthy AI-enabled systems, and for evaluating trustworthiness after system development. Although the principles behind MAST would seem more suitable for intelligence tasks given its focus on information quality and integrity, it is possible that these criteria may be applied to other human-in-the-loop systems used for information-processing and other human decision-making tasks. For example, AI-enabled systems in computer vision, natural language processing, and medical diagnostic tasks may all be rated according to the MAST criteria, including rating the system's sourcing (e.g., credibility of training data), or its ability to describe and propose alternative recommendations. Medical professionals and their patients may be more willing to trust an AI-derived diagnosis and treatment plan if the system was developed to include the MAST criteria of uncertainty, analysis of alternatives, and customer relevance.

It should be noted that several instruments have been developed to measure trust in automation, including instances of AI-enabled automation (Alsaid, Li, Chiou, & Lee, 2023; Kohn, de Visser, Wiese, Lee, & Shaw, 2021). Many of these instruments have been widely adopted, others have been independently validated. However, these instruments were mainly designed for research or technology evaluation purposes, rather than for technology development or operational settings. Therefore, although these instruments could be considered relatively robust when used appropriately, they suffer from similar limitations as the design frameworks and tools described previously. There remain wide translation gaps, and highly variable interpretation from principles to practice, given the hundreds of under-specified conditions and decisions that system designers and other practitioners face. For example, underlying many of these instruments is a nuanced presumption that assessing domain experts' trust in a particular technology, after they have experienced using the technology, could be some indication of the technology's trustworthiness. This presumed connection between trust and trustworthiness is then flattened in some practitioner circles, where high trust perceptions are equated with high technology trustworthiness, despite most trust experts being careful not to conflate the two.

To situate the MAST tool in the context of current trust scholarship, our primary objective is to assess the construct validity of MAST relative to human trust. Construct validity is the degree to which an instrument measures the construct it was designed to measure (Cronbach & Meehl, 1955). Approaches for evaluating construct validity include multivariate analytical tools, such as factor analysis (Raykov & Marcoulides, 2008; Tabachnick, Fidell, & Ullman, 2013), principal components analysis (PCA; Bandalos, 2018), and structural equation modeling (Kline, 2015). The goal of using multivariate analysis in construct validation is to capture, explain, and measure the amount of variation among items for a construct and to associate these with previously validated constructs (Chancey, Bliss, Yamani, & Handley, 2017; Jian, Bisantz, & Drury, 2000). This study aimed to validate MAST as an instrument for assessing trust by investigating how MAST items are associated with validated trust questionnaires.

3. General Method

To validate MAST in different contexts, we used the MAST checklist as a framework, in conjunction with our expertise and knowledge on state-of-the-art AI algorithms, to design two AI-DSS testbeds, one for identity-verification in a security screening task (Facewise) and another for text summarization and visualization in an investigative task to support intelligence reporting (READIT). We describe these two testbeds in more detail below.

3.1 Testbeds: Facewise and READIT

Facewise is a simulated 1-to-1 identity verification system that utilizes a pre-trained convolutional neural network, further fine-tuned for face recognition tasks using Cross-entropy loss. It compares an identification photo with a live or encounter photo and outputs a decision on whether they represent the same identity (match) or different identities (mismatch). Such AI-powered face-matching decision support systems are increasingly common at airport security checkpoints, such as CAT-C or CAT2 (Lim & Cantor, 2021).

READIT, which stands for the REporting Assistant for Defense and Intelligence Tasks, is an emulated natural language processing system that was designed to compile, summarize, and categorize documents of limited length (news articles, reports, microblogs) to expedite intelligence gathering and reporting. READIT first uses BERT (Devlin, Chang, Lee, & Toutanova, 2019) to generate outputs, after which we manually improved on the model outputs to enhance the usefulness and usability of the tool.

The case scenario for READIT was to assess MAST within the text summarization contexts that MAST was originally designed and evaluated for (Blasch et al., 2021). The identity verification scenario was selected to test the validity of the MAST checklist using a different type of AI capability, in a different type of task environment, while staying within a national security context subject to low risk tolerance. Our case scenario and AI-DSS testbeds were designed and developed based on information gathered from field visits to operational security screening environments, and bi-monthly consultations with operational stakeholders (i.e., national security researchers, practitioners, and analysts).

Both Facewise and READIT were developed using cloud-based services consisting of client-server model for user-AI interaction. In the Facewise system, we leveraged Amazon Web Services (AWS) and Google Cloud Platform (GCP) for efficient use of storage and resources. We built the client part of the platform with HTML5 and JavaScript. We collected the responses from participants on the client's side and sent them to the GCP through Python3 and Flask library to save them in the database. Similarly, the READIT system consisted of a JavaScript based client that enables the participant-AI interaction, and the server was built using Python3 and Flask library, hosted on GCP. Data visualizations on READIT were created to aid in better understanding of the dataset. The visualizations were implemented using D3.js, which is a popular open-source JavaScript library for creating custom interactive data visualizations. While participants were conducting the task, we logged system activities (e.g., button clicks, and relevant changes to the system state) to assess performance. The implementation code for READIT and Facewise is available at: https://github.com/nayoungkim94/PADTHAI-MM.

3.2 Constructs and Measures

For both DSS platforms, system features were manipulated to comprise two versions (High-MAST and Low-MAST) with eight outcome variables of interest: MAST criteria ratings; perceptions of risk, benefit, trust, credibility; task performance; self-reported engagement and usability. These constructs and measures are defined in more detail below.

Versions of the DSS: High-MAST and Low-MAST. System features refer to the available features that a DSS can provide its operators. Based on the MAST checklist, two levels of features for each platform were created: High-MAST and Low-MAST. High-MAST features were designed to score high ratings on each of the MAST criteria, resulting in a set of rich features that was supposed to be helpful to excel in the task. On the other hand, Low-MAST features were designed to score low ratings on each of the MAST criteria with a minimum set of necessary features included to be able to complete a task. In summary, the High-MAST versions could be described as providing more information about the DSS's performance given the task context, and the Low-MAST versions were designed to operate more like black-box systems. However, both High- and Low-MAST versions were designed to be as equal as possible in terms of engagement and usability. Appendices A and B delineate the MAST criteria and detailed feature descriptions for Facewise and READIT, respectively. More information about our development process and design decisions of our DSS testbeds are not the focus of this paper, but will be reported in detail in a forthcoming paper.

Variables of interest: MAST criteria, risk, benefit, trust, credibility, performance, engagement, and usability. Each DSS was evaluated based on descriptions of the MAST checklist and using a Likert-like scale of 1 to 4, with 1 being poor and 4 being excellent. Each MAST criterion was shown in a question format and accompanied by a corresponding feature description of the DSS. The MAST-total score was created by adding up these 9 criteria with a range of 9 to 36. Participant perception of risk and benefit was measured through two items derived from (Weber, Blais, & Betz, 2002). Risk was included due to the well-known relationship between trust and risk (Lee & See, 2004) and perceived benefit was included to check whether participants felt that using the DSS was beneficial for the task they were asked to complete. To measure trust, we used two commonly-used questionnaires. One is a previously validated, 12-item instrument known to measure general trust perceptions of automation (Jian et al., 2000; Spain, Bustamante, & Bliss, 2008). The second 15-item instrument measures trust by querying about specific types of information known to affect trust – purpose, process, and performance (Chancey et al., 2017). Because the MAST checklist largely focus on the presentation, availability, types, and quality of information presented by the AI system, we also included a measurement for message credibility (i.e., excluding source credibility), adopting a 3-item survey (Appelman & Sundar, 2016). Appendix C presents the scale, example items, number of items for each variable, and their Cronbach's alpha.

We measured performance through two variables, average task completion time and a scenario-specific performance metric. For Facewise, our scenario-specific performance metric focused on average accuracy on the identity verification task that took roughly 30 minutes to complete. The off-the-shelf performance of the algorithm used had an accuracy rate of 95% across test data during model training (Coşkun, Uçar, Yildirim, & Demir, 2017). However, with the database used in this experiment, its performance was around 60% for the difficult cases, and greater than 95% for the easy cases. Participants were not given this information about the algorithm's performance in advance nor were they informed about the potential difficulty levels of the cases in advance, but they were alerted to the fact that part of their task was to ensure that the correct decision was made with an algorithm that was potentially fallible. More information about the algorithm's performance could be accessed by participants only from the additional information option provided as part of the High-MAST interface.

For READIT, our scenario-specific performance metric was the score of a 250-word report completed within roughly 60 minutes. We asked participants to identify any present terrorist threat based on past news in a fictitious city named "Vastopolis" and to write a 250-word report detailing its aspects and reasons (e.g., type of terrorist activity, the group behind it, etc.) The task was designed based on the 2011 IEEE Visual Analytics Science and Technology (VAST) Challenge (IEEE SEMVAST Project, 2011). To score the reports, we created a rubric ranging from 1 to 5 based on similarity to the ground truth of the VAST challenge. A score of 1 indicated unsatisfactory content, which consisted solely of red herrings or non-ground truth clusters. A score of 2 represented less satisfactory content, mostly comprising red herrings or non-ground truth clusters. A score of 3 denoted satisfactory content, with more ground truth clusters than red herrings or non-ground truth clusters. A score of 4 signified mostly satisfactory content, primarily consisting of ground truth clusters. Finally, a score of 5 marked excellent content, comprising only ground truth clusters. The reports were color-coded for easier grading, using red for red herrings or non-ground truth clusters and bold for ground truth clusters. Two researchers independently rated participants' analytical reports. The grading scores largely converged (inter-rater reliability was 73.91%), and in case of any discrepancy among raters, the lower score was used. Apart from task performance, the other dependent variables and covariates were identical for Facewise and READIT.

To ensure that our implementation of different system features across the two testbeds would not cause major differences in perceived system usability and task engagement, and subsequently affect trust, risk, and benefit perceptions, we measured participants' perceived usability and engagement in the interactive task. Usability was assessed with a widely-used 10-item questionnaire known as the System Usability Scale (Brooke, 2020) and engagement was assessed with a 17-item questionnaire (Schaufeli, Salanova, Gonzalez-Roma, & Bakker, 2002). Appendix C reports more details of these measures. In addition, because study participants were experts (professionally trained in) face matching and intelligence analysis task domains, we manipulated task difficulty to ensure sufficient task engagement. For Facewise, we did this by hand-selecting 80 pairs of face images with known ground truth, with 40 of those image pairs representing easy tasks and the remaining 40 image pairs representing difficult tasks, presented in randomized order. Difficulty was defined from the perspective of the human operator, with difficult image pairs largely selected from a publicly available sibling database (Parkhi, Vedaldi, & Zisserman, 2015), and validated in a pilot study with a general population sample that was on average more likely to get the difficult pairs wrong. For READIT, we included "red herring" documents from the VAST Challenge dataset (IEEE SEMVAST Project, 2011) to encourage participant engagement in the task; several of these documents were related and collectively formed narratives that presented several plausible causes for the terrorist threat scenario. These "red herring" documents would then ideally cause sufficiently engaged participants to consider several highly plausible conclusions for their final report.

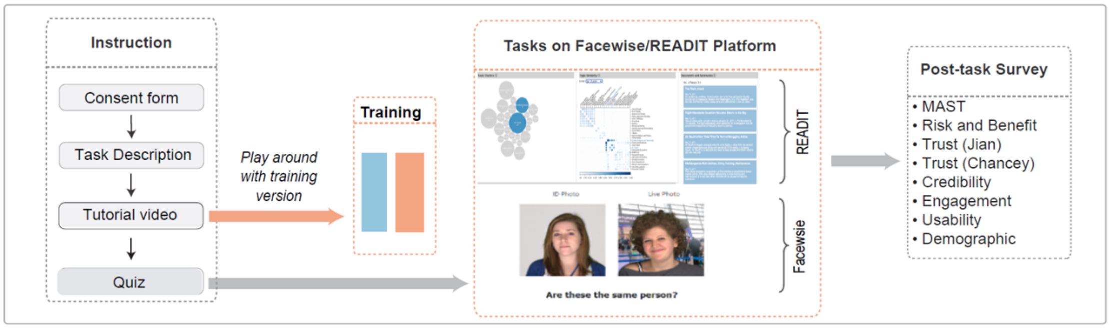


Figure 1: Study Procedure for Facewise and READIT.

3.3 Procedure

Figure 1 illustrates the general procedure of this study for Facewise and READIT. We first created a virtual hub using the web-based software platform Qualtrics (Qualtrics, 2020) for participants to access and complete the study. After random assignment to one of the two conditions (High-MAST or Low-MAST) by the researchers, participants were asked to input their given participant IDs on the first page of Qualtrics. Next, informed consent approved by our Institutional Review Board (IRB) was obtained. Then, participants were given a description of their task scenarios. To facilitate engagement and a sense of risk in the study scenario, in all conditions participants were told that they were being tasked to complete an important assignment, and that a previous agent assigned to their post failed in their respective tasks and was put on probation and subsequently demoted. After reading the task descriptions, participants then received a short training on their respective DSSs by watching a recorded video demonstration of the interface and features, and responding correctly to quiz questions about the video. All DSS versions were presented as technology aids that exist to supplement the participant's own abilities. Afterward, participants performed the study task using their respective DSS. Lastly, participants were asked to evaluate the system and their experience by responding to questionnaires including the MAST criteria, risk, benefit, trust, credibility, engagement, and usability. Given our targeted population of subject matter experts in national security, limited optional demographic information was collected to assess the representativeness of our sample population.

3.4 Data Analysis

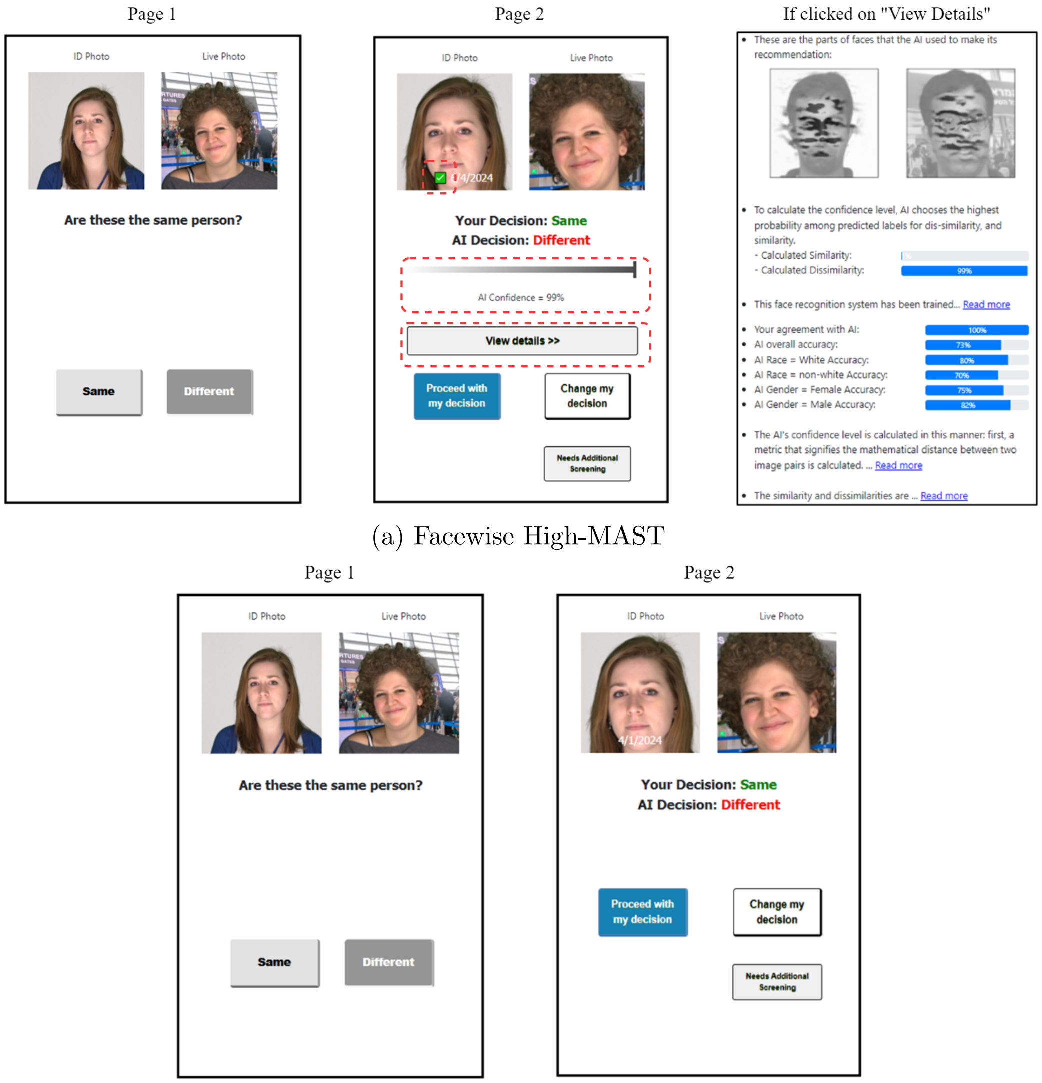
Data analysis was accomplished in JMP (SAS Institute Inc., 2023) and R using "dplyr" (Wickham, François, Henry, & Muller, 2019), "psych" (Revelle, 2018), "Rmisc" (Hope, 2013), and "compareGroups" (Subirana, Sanz, & Vila, 2014). Figures were created using "ggmap" (Kahle & Wickham, 2013) and (Auguie, 2017). To confirm associations between the MAST checklist and measures of trust, credibility, and other validated metrics, we performed the analysis in three steps. First, we explored if there are differences between the Low-MAST and High-MAST groups with respect to the dependent variables identified above. Secondly, linear associations between the individual metrics and the averaged MAST rating were separately established using simple linear regression. Multivariate analysis via principal components analysis (PCA) was then performed on the perceptual metrics for dimension reduction. Finally, we regressed the MAST ratings with the principal component (PC) scores. PCA was employed to find coherent and appropriate structures in the perceptual metrics within the first few principal components (Bandalos, 2018). To compare the different levels of Facewise and READIT, Analysis of Variance (ANOVA) was used. In addition, linear regression was used to further investigate the strength and directionality between MAST and other survey measures.

4. Experiment 1: Facewise

Participants in the Facewise experiment were told that they were airport security officers tasked to screen passengers by checking their identification materials with the assistance of Facewise, and they had roughly 30 minutes to complete a series of identification verification tasks which is roughly the length of an officer's shift in the document checker position (Greene, Kudrick, & Muse, 2014). Figure 2 outlines the similarities and differences between the two levels of Facewise: High-MAST and Low-MAST. For both levels, Page 1 asks for an initial judgment of human operators. We adopted this structure based on previous work, which we found would increase accuracy (Salehi, Chiou, Mancenido, Mosallanezhad, Cohen, & Shah, 2021). In Page 1 (Figure 2), the left image with an off-white background presents the ID photo and the right image with an airport background provides a "live" photo, supposedly taken at the airport. For both levels, these images were cropped and zoomed in for Page 2, which is where most of the differences between High-MAST and Low-MAST appear. Three red dotted lines highlight these differences including the Crossmark/Checkmarks, AI confidence, and a "View Details" button. For more details regarding the AI-DSS features and how they map to each of the MAST criteria, please refer to Appendix A.

4.1 Participants

A total of 152 subject matter experts, U.S. Transportation Security Officers (TSOs), were recruited from three major U.S. airports in Arizona, Nevada, and California, split across 11 days of data collection at the participating airports. Six participants were removed due to very high response time and low accuracy, resulting in 73 participants each for the High-MAST and Low-MAST conditions. On average, participants spent 76 minutes to complete the entire study, including onboarding and responses to questionnaire items. Because participants were federal employees, we were not permitted to provide compensation despite their participation being voluntary and outside of their regular duties. Therefore, light refreshments were provided to appreciate their participation in the study. Because all participants were volunteers and not required to participate, we assumed they were sufficiently motivated by our study objectives to complete this study to the best of their ability. Table 5 in Appendix F reports the available participant demographics across the Facewise conditions. Race, ethnicity, and gender items were not collected or reported due to expressed concerns by some of our collaborative partners, given our limited population of subject matter experts, and the sensitive nature of this information.



Are these the same person?

Your Decision: Same
AI Decision: Different

- These are the parts of faces that the AI used to make its recommendation:
- To calculate the confidence level, AI chooses the highest probability among predicted labels for dis-similarity, and similarity.
  - Calculated Similarity: 1%
  - Calculated Dissimilarity: 99%
- This face recognition system has been trained... Read more
- Your agreement with AI: 100%
- AI overall accuracy: 73%
- AI Race = White Accuracy: 80%
- AI Race = non-white Accuracy: 70%
- AI Gender = Female Accuracy: 75%
- AI Gender = Male Accuracy: 82%
- The AI's confidence level is calculated in this manner: first, a metric that signifies the mathematical distance between two image pairs is calculated. ... Read more
- The similarity and dissimilarities are ... Read more

(a) Facewise High-MAST

Are these the same person?

Your Decision: Same
AI Decision: Different

(b) Facewise Low-MAST

Figure 2: Facewise High-MAST (top) and Facewise Low-MAST (down). Red dashes highlight the differences between Low and High platforms. Compared to the Facewise Low-MAST, High-MAST version has more interactive features including the ID expiration check, AI confidence level, and "View Details" page.

4.2 Results and Discussion

Table 1 reports descriptive statistics including mean (M) and standard deviations (SD) for the study variables. Results of F statistics in Figure 3 (a) show that participants in the High-MAST group rated Facewise higher across all nine MAST criteria, and this difference was significant. The higher trust ratings in High-MAST compared to Low-MAST was also significantly different for the Jian et al., 2000 score. The High-MAST group found Facewise less risky than the Low-MAST group. Moreover, the High-MAST group rated Facewise more beneficial than those in the Low-MAST group. No significant difference in credibility ratings was found between the two conditions, possibly due to the similar presence of errors by Facewise in both levels. While the High-MAST group spent significantly more time on the task than the Low-MAST group, performance was not significantly different between the two groups. The High-MAST group made slightly more accurate decisions than Low-MAST, but this difference was not significant. No significant differences in engagement and usability were found between the High-MAST group and Low-MAST group, supporting our goal of designing both versions to be relatively equal in terms of levels of engagement and perceived usability.

Regression analysis shows that MAST ratings are positively associated with trust ratings; people who rated trust highly also tended to rate MAST highly. Increasing the MAST score by 1 would increase one of the trust scores (Jian et al., 2000) by 0.1 ($F(1, 144) = 64.94$, $p < .001$, $\beta = 0.1$, $R2 = 0.31$) and the other trust score (Chancey et al., 2017) by 0.12 ($F(1, 144) = 87.83$, $p < .001$, $\beta = 0.12$, $R2 = 0.37$). A positive relationship between MAST and credibility scores was also found; increasing the MAST score by 1 would increase credibility by 0.11 ($F(1, 144) = 62.96$, $p < .001$, $\beta = 0.11$, $R2 = 0.30$). Figure 4 shows the regression plots. Furthermore, this study found that there was a negative correlation between MAST and risk score; increasing the risk score by 1 would decrease the MAST score by 2.2 ($F(1, 144) = 24.32$, $p < .001$, $\beta = -2.2$, $R2 = 0.14$). Finally, there was a positive relationship between the MAST and benefit score; increasing the benefit score by 1 would increase the MAST rating by 3.7 ($F(1, 144) = 71.89$, $p < .001$, $\beta = 3.7$, $R2 = 0.33$).

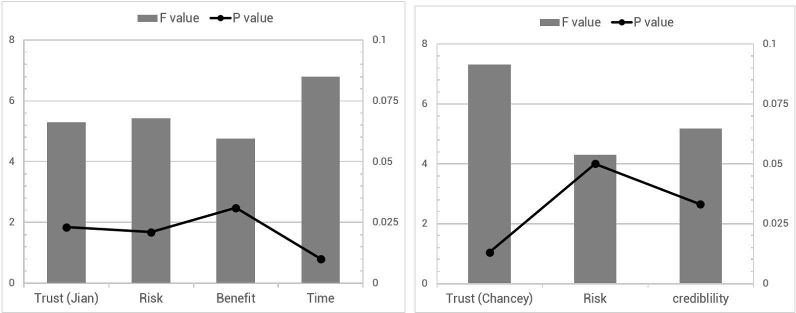
To further validate the association between MAST and other study variables, we needed to run multiple regression analysis. However, because trust (Jian et al., 2000; Chancey et al., 2017), risk, benefit, and credibility were highly correlated, it is inappropriate to run multiple regression analyses. Therefore, we applied PCA to reduce the dimensionality within our dataset. The result of PCA shows that the first two principal components explain 84.06% of variation within the dataset. The first principal component can be perceived as an overall average of trust, risk, benefit, and credibility, while the second principal component is mainly related to negative perceptions about risk. These two principal components were used as new variables for a linear regression analysis with MAST performed for each level, High- and Low-MAST. We found that MAST-total (aggregating all MAST criteria) was highly associated with the first principal components ($F(1, 144) = 100.92$, $p < .001$, $\beta = 0.19$, $R2 = 0.41$). Figure 5 provides additional details about PCA and regression results.

5. Experiment 2: READIT

READIT participants were told they were intelligence analysts in a fictional major city in the United States who are tasked with monitoring the news for any ongoing threats to public safety. READIT participants were given a specific assignment to use READIT to quickly locate and search through relevant news articles and uncover a terrorist activity that had gone unnoticed for the previous five months. Figure 6 illustrates the similarities and differences between the High-MAST and Low-MAST levels of READIT. Four red dotted lines highlight the differences including the availability of “documents” and “about” tabs (Appendix D), the “topic clusters” bubble graph, the sorting option by cluster relationship

	Facewise			**READIT**		
	High	**Low**	*p*	**High**	**Low**	*p*
MAST-total	26.5 (5.73)	21.0 (5.77)	< 0.001*	27.8 (5.38)	19.9 (5.14)	0.002*
1. Sourcing	2.85 (0.84)	2.48 (0.90)	0.011*	3.27 (0.47)	2.50 (1.17)	0.05*
2. Uncertainty	2.85 (0.83)	2.32 (0.97)	< 0.001*	2.91 (0.70)	2.42 (0.79)	0.129
3. Distinguishing	3.19 (0.78)	2.15 (0.88)	< 0.001*	3.27 (0.65)	2.25 (0.87)	0.004*
4. Alternatives	2.79 (0.82)	2.33 (0.91)	< 0.001*	2.91 (0.94)	1.25 (0.45)	0.001*
5. Relevance	3.00 (0.69)	2.34 (0.89)	< 0.001*	3.18 (0.75)	3.17 (0.72)	0.961
6. Logic	2.97 (0.87)	2.07 (0.96)	< 0.001*	3.18 (0.87)	2.25 (0.97)	0.024*
7. Change	2.88 (0.83)	2.51 (0.82)	0.008*	2.82 (0.75)	1.75 (0.97)	0.007*
8. Accuracy	2.82 (0.87)	2.42 (0.82)	0.005*	3.18 (0.75)	1.92 (0.67)	< 0.001*
9. Visualization	3.14 (0.75)	2.42 (0.86)	< 0.001*	3.09 (0.94)	2.42 (0.90)	0.095
Trust (Jian)	4.62 (1.12)	4.18 (1.15)	0.023*	5.11 (0.95)	4.42 (1.08)	0.119
Trust (Chancey)	4.26 (1.28)	4.00 (1.16)	0.188	4.57 (1.25)	3.73 (1.35)	0.135
Chancey (Performance)	4.24 (1.42)	3.95 (1.28)	0.188	4.85 (1.31)	3.93 (1.47)	0.126
Chancey (Process)	4.68 (1.39)	4.52 (1.36)	0.472	4.76 (1.56)	4.48 (1.53)	0.669
Chancey (Purpose)	3.87 (1.28)	3.53 (1.16)	0.093	4.09 (1.15)	2.77 (1.20)	0.013*
Risk	2.67 (1.11)	3.10 (1.09)	0.021*	2.55 (0.93)	3.33 (0.89)	0.05*
Benefit	3.37 (0.99)	3.01 (0.98)	0.031*	3.45 (0.93)	3.00 (1.13)	0.303
Credibility	4.31 (1.24)	4.23 (1.29)	0.712	5.39 (0.96)	4.44 (1.03)	0.033*
Average response time (seconds)	13.3 (4.87)	11.3 (4.46)	0.010*	274 (90.7)	214 (93.6)	0.137
Performance (in Platforms)	0.77 (0.08)	0.75 (0.07)	0.097 (accuracy)	2.82 (1.94)	3.33 (1.67)	0.505 (report)
Engagement	3.96 (1.07)	3.99 (1.10)	0.874	4.37 (0.78)	4.52 (1.28)	0.736
Usability	3.58 (0.70)	3.65 (0.60)	0.494	3.49 (0.94)	3.90 (0.67)	0.248

Table 1: Means, Standard Deviation (in parentheses), and *p*-values of study variables for High-MAST and Low-MAST groups across Facewise and READIT platforms. Asterisk(*) emphasizes the significant differences.



(a) Facewise (b) READIT

Figure 3: The *F*-test results and corresponding *p*-values for significant variables in (a) Facewise and (b) READIT are displayed as grey bars and black dots, respectively.

strength, and the complete news pieces. For more details regarding the AI-DSS features and how they map to each of the MAST criteria, please refer to Appendix B.

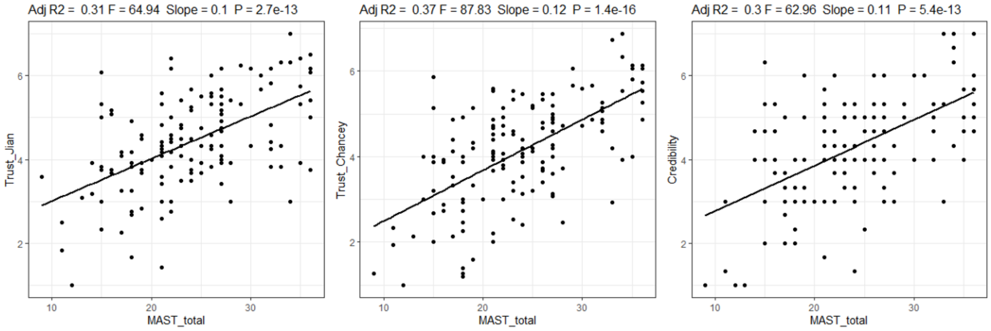
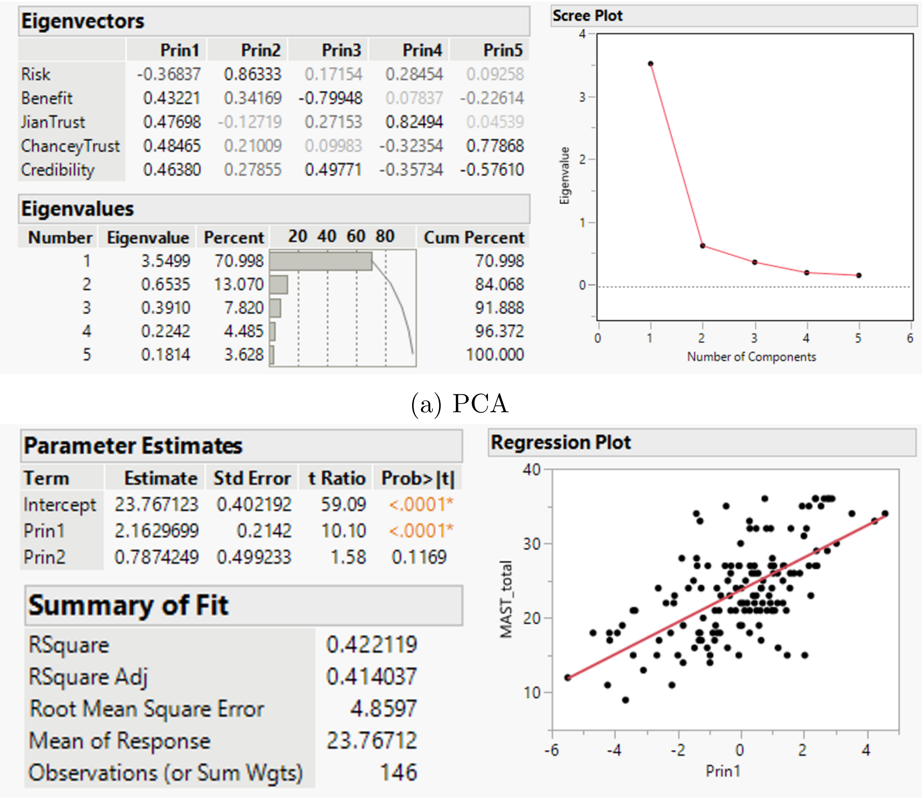


Figure 4: Least Squares Regression plots for Facewise.



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Risk	-0.36837	0.86333	0.17154	0.28454	0.09258
Benefit	0.43221	0.34169	-0.79948	0.07837	-0.22614
JianTrust	0.47698	-0.12719	0.27153	0.82494	0.04539
ChanceyTrust	0.48465	0.21009	0.09983	-0.32354	0.77868
Credibility	0.46380	0.27855	0.49771	-0.35734	-0.57610

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.5499	70.998		70.998
2	0.6535	13.070		84.068
3	0.3910	7.820		91.888
4	0.2242	4.485		96.372
5	0.1814	3.628		100.000

(a) PCA

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	23.767123	0.402192	59.09	<.0001*
Prin1	2.1629699	0.2142	10.10	<.0001*
Prin2	0.7874249	0.499233	1.58	0.1169

Summary of Fit

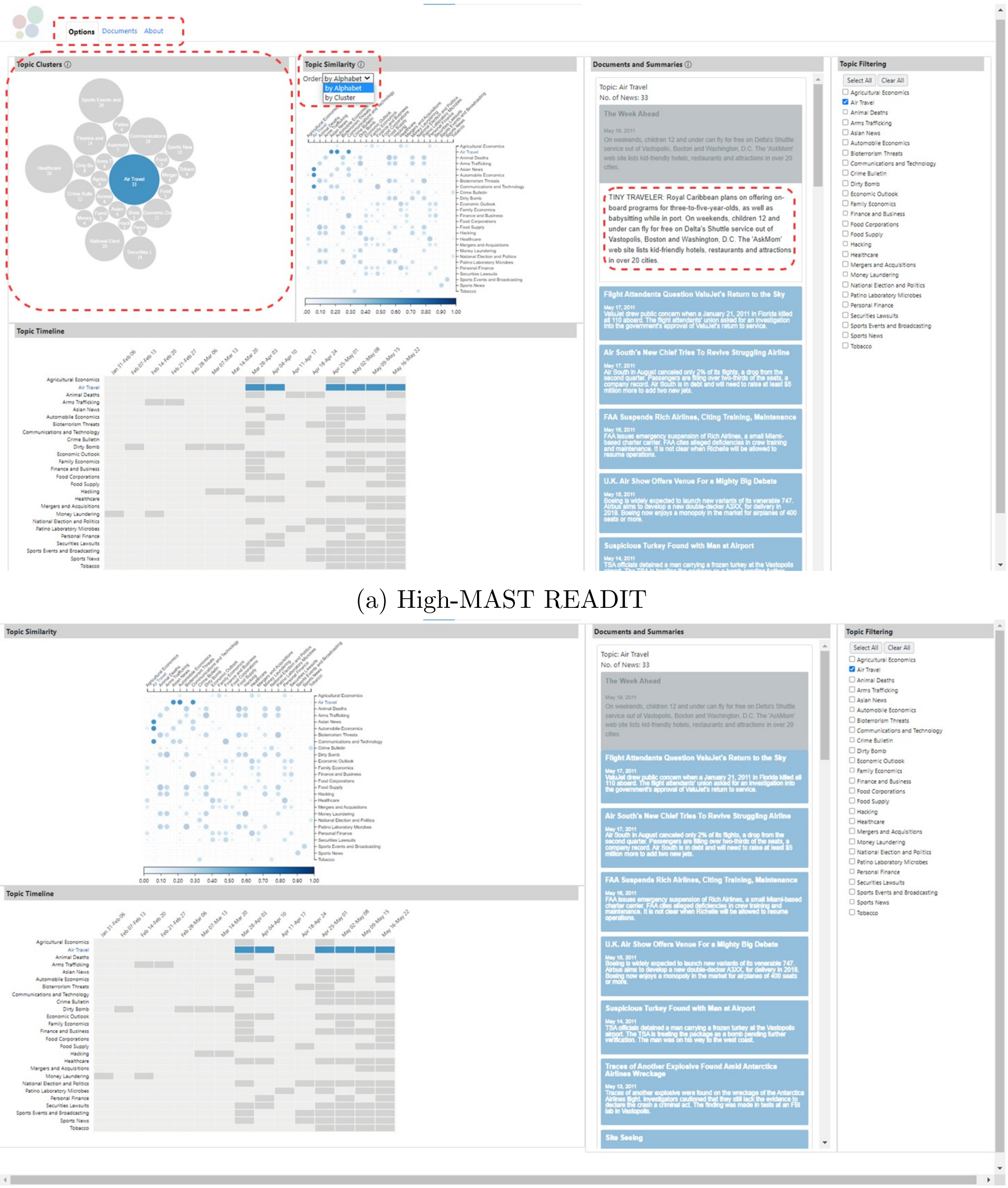
RSquare	0.422119
RSquare Adj	0.414037
Root Mean Square Error	4.8597
Mean of Response	23.76712
Observations (or Sum Wgts)	146

(b) Linear Regression

Figure 5: PCA (top) and Linear regression (down) results for Facewise.

5.1 Participants

A total of 25 Intelligence Analysts (IAs) from the U.S. Department of Homeland Security (DHS) were recruited to complete our study, administered through Microsoft Teams or Zoom, over a period of 19 days. Two participants were unable to complete the study due to unexpected scheduling restrictions. The resulting High-MAST and Low-MAST versions of READIT were tested with a sample of 11 and 12 IAs, respectively. On average, participants



Documents and Summaries ⓘ

Topic: Air Travel
No. of News: 33

The Week Ahead
May 19, 2011
On weekends, children 12 and under can fly for free on Delta's Shuttle service out of Vastopolis, Boston and Washington, D.C. The 'AskMom' web site lists kid-friendly hotels, restaurants and attractions in over 20 cities.

TINY TRAVELER: Royal Caribbean plans on offering on-board programs for three-to-five-year-olds, as well as babysitting while in port. On weekends, children 12 and under can fly for free on Delta's Shuttle service out of Vastopolis, Boston and Washington, D.C. The 'AskMom' web site lists kid-friendly hotels, restaurants and attractions in over 20 cities.

Flight Attendants Question ValuJet's Return to the Sky
May 17, 2011
ValuJet drew public concern when a January 21, 2011 in Florida killed all 110 aboard. The flight attendants' union asked for an investigation into the government's approval of ValuJet's return to service.

Air South's New Chief Tries To Revive Struggling Airline
May 17, 2011
Air South in August canceled only 2% of its flights, a drop from the second quarter. Passengers are filling over two-thirds of the seats, a company record. Air South is in debt and will need to raise at least $5 million more to add two new jets.

FAA Suspends Rich Airlines, Citing Training, Maintenance
May 16, 2011
FAA issues emergency suspension of Rich Airlines, a small Miami-based charter carrier. FAA cites alleged deficiencies in crew training and maintenance. It is not clear when Richelle will be allowed to resume operations.

U.K. Air Show Offers Venue For a Mighty Big Debate
May 15, 2011
Boeing is widely expected to launch new variants of its venerable 747. Airbus aims to develop a new double-decker A3XX, for delivery in 2018. Boeing now enjoys a monopoly in the market for airplanes of 400 seats or more.

Suspicious Turkey Found with Man at Airport
May 14, 2011
TSA officials detained a man carrying a frozen turkey at the Vastopolis airport. The TSA is treating the package as a bomb pending further verification. The man was on his way to the west coast.

(a) High-MAST READIT

Traces of Another Explosive Found Amid Antarctica Airlines Wreckage
May 13, 2011
Traces of another explosive were found on the wreckage of the Antarctica Airlines flight. Investigators cautioned that they still lack the evidence to declare the crash a criminal act. The finding was made in tests at an FBI lab in Vastopolis.

Site Seeing

(b) Low-MAST READIT

Figure 6: High-MAST READIT (top) and Low-MAST READIT (down). Compared with the Low-MAST READIT, High-MAST READIT has more interactive features (Topic Clusters, Topic Similarity, original documents, clickable timelines, etc.) to demonstrate the MAST criterion.

spent 75 minutes to complete the study, including onboarding and responses to questionnaire items. We were not permitted to compensate participants monetarily because they were federal employees. However because participants were self-selected volunteers who responded to our recruitment script and were willing to spend time completing our study, we assumed

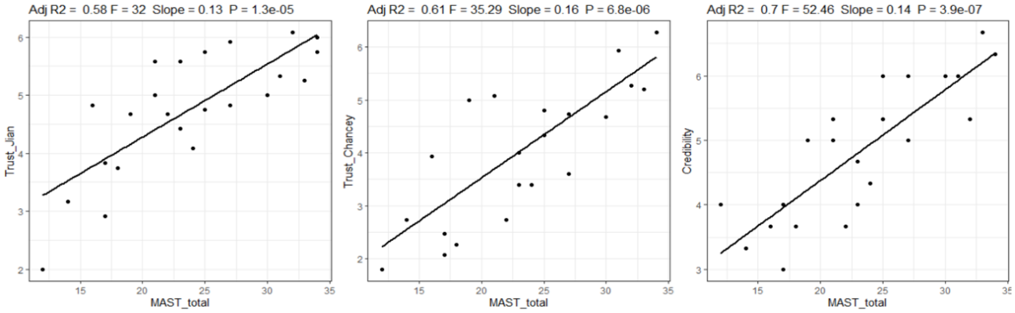


Figure 7: Least Squares Regression plots for READIT.

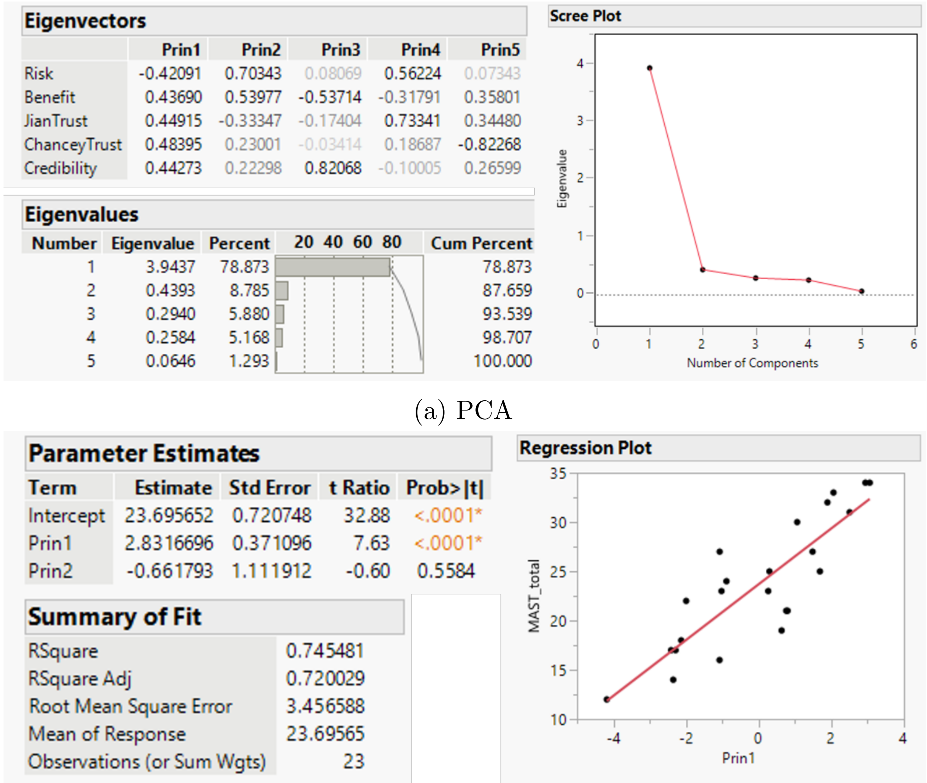
they were sufficiently motivated to complete this study to the best of their ability. Table 6 in Appendix F reports the participant demographics per condition.

5.2 Results and Discussion

Table 1 reports descriptive statistics including mean and standard deviations for the study variables. Results show that the High-MAST group rated READIT higher on the MAST checklist than the Low-MAST group, and this was significantly different for six out of nine MAST criteria (i.e., except for uncertainty, relevance, and visualization). Trust ratings were also generally higher for those in the High-MAST group; however, the difference is only significant for the "purpose" dimension of the Chancey et al., 2017 trust score. Moreover, the High-MAST group compared to the Low-MAST group found READIT less risky to use, and more credible. No significant differences in performance were found between High-MAST and Low-MAST groups in terms of average response time, or on their 250-word report, although descriptively the Low-MAST group spent less time completing the task and had higher performance scores than the High-MAST group. No significant differences were found in engagement and usability ratings between High-MAST and Low-MAST groups, supporting our intent to keep the different READIT versions roughly equivalent in terms of levels of engagement and usability. Figure 3 (b) shows the F values of significant variables in READIT.

Regression analysis showed that MAST ratings are positively associated with trust ratings. There is a positive relationship between MAST and both the trust scores (Jian et al., 2000; Chancey et al., 2017); increasing MAST by 1 increases one of the trust scores (Jian et al., 2000) by 0.13 ($F(1, 21) = 32$, $p < .001$, $\beta = 0.13$, $R2 = 0.58$) and increases the other trust score (Chancey et al., 2017) by 0.16 ($F(1, 21) = 35.29$, $p < .001$, $\beta = 0.16$, $R2 = 0.61$). A positive relationship was also found between MAST and credibility scores; increasing the MAST score by 1 would increase credibility ratings by 0.14 ($F(1, 21) = 52.46$, $p < .001$, $\beta = 0.14$, $R2 = 0.70$). Figure 7 shows the regression plots. This study also found that there was a negative relationship between MAST and risk score; increasing the MAST score by 1 would decrease perceived risk by 4.9 ($F(1, 21) = 24.89$, $p < .001$, $\beta = -4.9$, $R2 = 0.52$). In addition, there was a positive relationship between MAST and perceived benefit; increasing the benefit score by 1 would increase MAST by 4.2 ($F(1, 21) = 17.19$, $p < .001$, $\beta = 4.2$, $R2 = 0.42$).

Because trust (Jian et al., 2000; Chancey et al., 2017), risk, benefit, and credibility were highly correlated for the READIT platform, we did not run multiple regression and instead used PCA. The PCA results show that the first two principal components can explain 87.66% of variation within the dataset. The first principal component can be interpreted as an overall average of trust, risk, benefit, and credibility. However, the second principal component was primarily related to negative perceptions about risk. For all observations, two PC scores were calculated using each principal component and these scores served as the regressors for further analysis. We found that averaging across all MAST criteria to produce a MAST-total score can significantly predict the first principal components ($F(1, 144) = 60.07$, $p < .001$, $\beta = 0.26$, $R2 = 0.74$). Figure 8 provides more details about the PCA and regression results.



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Risk	-0.42091	0.70343	0.08069	0.56224	0.07343
Benefit	0.43690	0.53977	-0.53714	-0.31791	0.35801
JianTrust	0.44915	-0.33347	-0.17404	0.73341	0.34480
ChanceyTrust	0.48395	0.23001	-0.03414	0.18687	-0.82268
Credibility	0.44273	0.22298	0.82068	-0.10005	0.26599

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.9437	78.873		78.873
2	0.4393	8.785		87.659
3	0.2940	5.880		93.539
4	0.2584	5.168		98.707
5	0.0646	1.293		100.000

(a) PCA

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	23.695652	0.720748	32.88	<.0001*
Prin1	2.8316696	0.371096	7.63	<.0001*
Prin2	-0.661793	1.111912	-0.60	0.5584

Summary of Fit

RSquare	0.745481
RSquare Adj	0.720029
Root Mean Square Error	3.456588
Mean of Response	23.69565
Observations (or Sum Wgts)	23

(b) Linear Regression

Figure 8: PCA (top) and Linear regression (down) results for READIT.

6. General Discussion

In this study, we recruited subject matter experts to interact with an AI-DSS in their field, either Facewise or READIT, and tested two different levels of each DSS, a High-MAST version or a Low-MAST version. We discuss our findings with respect to the experts' ratings of these systems using MAST, validated trust questionnaires, other perception metrics, and joint system performance. Finally, we elaborate on our analysis of the MAST items and participant perceptions, and conclude with some caveats regarding our study approach and findings.

Overall MAST ratings. The application of MAST to both Facewise and READIT resulted in notable differences in MAST ratings. Under High-MAST conditions, Facewise achieved higher scores across all criteria (9/9), while READIT achieve higher scores on 6 out of the 9 criteria. This difference between Facewise and READIT indicates that the type of system and use context matters when applying the MAST checklist. For an image processing and signal detection type outputs like Facewise, using MAST to evaluate elements like accuracy, source reliability, and user interface clarity may be more straightforward, as reflected in the consistently higher ratings across all criteria in the High-MAST condition. In contrast, READIT as a text summarization system is riddled with the complexities of natural language processing model outputs and the semiotics of text interpretation. For example, our team was particularly challenged in designing appropriate visualizations for model outputs and explanations that could distinguish between High-MAST and Low-MAST READIT systems. In the end, the MAST ratings for the *Visualization* criterion were not significantly different. Moreover, the lack of statistical significance between the *Uncertainty* and *Customer Relevance* criteria may also point to High-MAST design features that had marginal to no impact. In the Low-MAST system, uncertainty measures such as tf-idf and cosine similarity scores (see Appendix B) were omitted, unlike in the High-MAST system where they were included. The results could suggest that these uncertainty scores were either ineffective in conveying uncertainty or were implemented in a manner that limited their usefulness. The same observation applies to the feature of filtering by location and topic, which was incorporated specifically to meet the *Customer Relevance* criterion, and was a distinguishing feature between the two system versions. The rating similarities between the systems could indicate that these features did not significantly influence users' perceptions of *Customer Relevance*.

Perceptions of trust. Evaluation of the Facewise systems revealed significantly different trust ratings (as per Jian et al., 2000), with higher ratings observed for the High-MAST condition. In contrast, the higher trust ratings observed in the High-MAST READIT system compared to its Low-MAST counterpart were not statistically different. This discrepancy may have stemmed from the smaller participant pool evaluating the READIT system, leading to increased standard errors in the observed differences. It was a challenge to recruit for READIT, given the relative inaccessibility of remotely-recruited working intelligence analysts relative to the on-site recruited Transportation Security Officers, and the general challenge of recruiting subject matter experts to volunteer participation in research studies. Additionally, for the trust constructs in the Chancey et al., 2017 questionnaire, the Facewise systems did not exhibit significant differences, despite the Jian et al., 2000 ratings demonstrating otherwise. We speculate that this might be due to the nature of the Jian et al., 2000 questionnaire items being valenced both negatively and positively whereas for the Chancey et al., 2017 questionnaire, the items are all positively valenced. Prior research has shown that item valence can affect responses in trust measures (Gutzwiller, Chiou, Craig, Lewis, Lematta, & Hsiung, 2019), and that negatively valenced trust items may result in more variable responses compared to positively valenced trust items (Schroeder, Chiou, & Craig, 2021). More research is needed to investigate why these two instruments measuring the same construct could result in different responses (e.g., (Long, Sato, Millner, Loranger, Mirabelli, Xu, & Yamani, 2020)).

Despite being evaluated by a considerably smaller group of participants, the READIT systems demonstrated a notable difference in the Chancey et al., 2017 trust measure, but only on the Purpose dimension. No other differences were evident in other trust measurements. It is important to note that the "purpose" items in Chancey et al., 2017's measure are designed to assess participants' belief in the READIT system's ability to assist them in their tasks or missions, even amid uncertainties or perceived errors. This difference in ratings could therefore be linked to the inherent challenges and semiotic nature of text summarization tasks, as opposed to the more straightforward outcome of face recognition tasks. The higher rating for the "purpose" dimension in the High-MAST READIT group suggests that users valued the system and its associated features as a helpful tool in summarizing massive amounts of documents.

Other user perception metrics. Comparative analysis of other user perception metrics revealed that high-MAST versions of the decision support systems generally led to reduced perceived risk and enhanced perceptions of benefit and credibility, aligning with our initial hypotheses. However, for READIT the "benefit" ratings were not significantly different between versions, and for Facewise, "credibility" ratings were not significantly different between versions. Notably, these items exhibited greater variability, indicating divergent perceptions of Facewise's credibility, likely influenced by the model's errors in tasks that may have been easy for our expert participants. In READIT's Low-MAST group, the "benefit" ratings varied more, suggesting that some of the analysts who participated in the study still found the system useful with respect to the imagined increased task load that would have come from manually inspecting the raw document data on their own.

In terms of "engagement" and "usability", we observed no significant differences between the High- and Low-MAST versions for both platforms. This uniformity in perceived engagement and usability ratings indicates a consistent user experience. Although this outcome was not a deliberate objective when we designed the platforms, we were careful to maintain this consistency in the low and high-MAST versions to minimize the impact of usability or ease of use on the evaluation of the systems.

Joint system performance. Neither platform showed significant differences in system performance between the High and Low-MAST versions, whether in face-matching accuracy for Facewise or the report score for READIT. This aligns with prior research suggesting that AI transparency or trustworthiness *alone* does not necessarily result in improved human performance (Schelble, Lancaster, Duan, Mallick, McNeese, & Lopez, 2023; Palanski & Yammarino, 2011). The absence of observed differences in our study could also be attributed to factors such as limited variation in the image database for Facewise, and the use of under-optimized AI algorithms common to both versions. In the case of READIT, despite clarification in the participant onboarding video, there might still have been some confusion among participants on how to interpret the bubble graph topic clusters. We discovered this issue during pilot tests with non-expert participants, who mistook the size of the bubble graph topic clusters for importance rather than topic frequency in the anomaly detection task. We attempted to correct this in our onboarding video by clarifying how to interpret the bubble graph, but ultimately we did not test to confirm their understanding or use of the bubble graph, which was one of the more salient differences between the High- and

Low-MAST versions of READIT. Designing the READIT interface was challenging due to project time constraints and the need to balance the presentation of detailed information with ease of navigation on a single browser page, without overly guiding participants to the correct answers. Future research could further refine these AI-DSS testbeds, improving AI performance, task variation, and optimizing the level of information detail in the interfaces.

Association between MAST and user perception metrics. Principal Components Analysis (PCA) and Principal Components Regression (PCR) were conducted to assess whether MAST could accurately capture key constructs from well-established user perception metrics. These metrics, which show significant marginal associations with MAST ratings, include trust measures from Jian et al., 2000 and Chancey et al., 2017, and measures of benefit, credibility, and risk. The primary goal of PCA in this context was to produce comprehensive summaries that capture the majority of variation within these metrics. The analysis showed that for both platforms, the first two eigenvalues accounted for over 80% of the total, suggesting that the first two principal components are sufficient in explaining the majority of variation in the data.

In both Facewise and READIT, the first principal component (PC) uniformly displayed positive loadings for all metrics except risk. This consistent pattern across both platforms indicates that a uniformly weighted average of these metrics, negatively weighted for risk, effectively captures the essential constructs of user perceptions in the evaluated technologies. Conversely, the second PC showed significant positive loadings exclusively for risk and benefit. This pattern suggests that participants tend to conduct a risk-benefit analysis, with the pronounced loading on risk indicating a stronger focus on risk assessment when evaluating AI systems.

The regression analysis of the first two PC scores against the MAST ratings revealed significant associations with the scores of the first PC, but not with those of the second. This finding indicates that the MAST ratings predominantly align with the factors captured by the first PC. Since the first PC primarily reflects a uniformly weighted combination of the user perception metrics, with an inverse weighting for risk, it can be inferred that MAST ratings are similar to an averaged rating of these metrics. This suggests that MAST effectively captures a broad spectrum of user perceptions, particularly trust, benefit, and credibility, while inversely accounting for risk. However, the lack of association with the second PC, which focuses more on risk-benefit analysis, implies that MAST may not fully capture the nuances of how users weigh risks against benefits when evaluating AI-enabled technologies such as Facewise and READIT. These potential nuances would further speak to the challenge of soliciting input on some of the MAST criteria, input that may vary widely depending on the experience level and perspectives of the responding stakeholders (Ananny & Crawford, 2016).

Study limitations. The AI-DSSs in this study were intentionally designed to align with either High-MAST or Low-MAST ratings. This methodology might invite criticism because the MAST tool, which required validation, was also employed in designing the experimental manipulations. However, we assert the validity of this approach based on the independence of the raters (i.e., the recruited study participants). The participants evaluating the technologies were not involved in the design process, nor the development of MAST, ensuring that their ratings were independent of self-serving biases. Further, intentionally aligning the designed features with the MAST checklist was necessary for the internal validation of the tool. This designed distinction allowed us to assess whether MAST, as an evaluative tool, could effectively differentiate between technologies with varying MAST alignment levels. Establishing internal validity serves as a foundation toward a broader collaborative effort to validate the tool externally. Additionally, this study sets a benchmark for future applications of MAST as a tool across various technological domains and ecological contexts.

Although MAST ratings were highly associated with trust, our results do not factor in whether trust or distrust levels were calibrated with system performance. Such an analysis may be possible for Facewise, in which system reliability can be precisely gauged using signal detection metrics. In contrast, trust and distrust calibration is difficult to precisely define for the READIT platform because it does not offer direct recommendations or answers that could be rated as easily. Future studies should consider these different forms of decision support, and how those different forms can affect trust responses (Chiou & Lee, 2023). Finally, although the signals were strong for READIT, we could not reach our intended sample size within the timeline we had, due to the challenge of recruiting intelligence analysts. Lastly, this study was focused on the use and validation of MAST specifically; a comprehensive review and comparison of MAST against other similar frameworks would be a valuable exercise, but beyond the scope of this project. Other literature has reviewed similar tools for trust assessment (Kohn et al., 2021; Alsaid et al., 2023), and MAST might be used in conjunction with some of these other tools alongside a work-centered field-based approach (Roth, Bisantz, Wang, Kim, & Hettinger, 2021) to achieve a more comprehensively designed and functionally trustworthy system.

7. Conclusion and Future Directions

The primary objective of this study was to establish the utility of the Multisource AI Scorecard Table (MAST) for evaluating the trustworthiness of AI-enabled decision support systems (AI-DSSs). This resulted in an interesting opportunity to evaluate whether the tradecraft standards behind MAST are related to the existing tools developed by the scientific community of trust researchers. The results of our study show strong associations between MAST and trust assessments across two domains of application. While MAST was initially conceptualized for intelligence reporting tools like READIT, by testing two AI-DSSs developed for different mission critical tasks, and with field experts, we also demonstrated the utility of MAST across high-consequence domains, and that these patterns of associations persist with automated identity verification systems as well.

Compared to other frameworks for designing or evaluating AI-DSSs, a benefit of MAST is that it is derived from and operationalized by a practitioner community (Blasch et al., 2021). Thus, the underlying principles of MAST are more likely to be “customer relevant” and accepted by the Intelligence Community, while aligning with an empirical and scholarly understanding of trust and credibility that we report here. However, just as high quality analytical reporting may not translate into good decision making, it is important to note that high MAST ratings do not necessarily translate to improved performance of a human-AI decision system, given the variety of factors that can contribute to this performance, including factors in the task environment, cognitive workload (Sargent, Walters, & Wickens, 2023), available system features, and task difficulty. Furthermore, it is still possible that high *intended* MAST ratings by a design team may not translate to actual perceptions and subsequently higher ratings, or actual benefit by a user of the system. Additional testing should be done with other AI systems, including key factors in the organizational and task environment (Chiou & Lee, 2023), and more formal risk analyses. In-depth exploration of the behavioral data captured during task performance may also shed light on the gap between trust perceptions and trustworthiness.

Acknowledgments

This material is based on work supported by the U. S. Department of Homeland Security under Grant Award Number 17STQAC00001-05-00. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Homeland Security. PS contributed the initial complete draft of the paper. PS, YB, MM contributed to data analysis. NK, MC, YB, SB contributed written sections and response to reviewers. PS, AP, AM, NK, YW, JZ contributed to design and development of Facewise and READIT. PS, AP, and MC contributed other study materials. PS, AP, MC, YB, NK, AM contributed to data collection. MM, JS, EB contributed to study conception, design, interpretation, and participant recruitment. EC contributed to all aspects of the study. All authors reviewed and approved the final version of this paper.

References

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, *14*, 1192020.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media society*, 1–17.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism and Mass Communication Quarterly*, *93*(1), 59–79.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, New York, NY.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4–1.

Blasch, E., Sung, J., & Nguyen, T. (2021). Multisource AI scorecard table for system evaluation.. Presented at AAAI FSS-20: Artificial Intelligence in Government and Public Sector, Washington, DC, USA.

Blasch, E., Sung, J., Nguyen, T., Daniel, C. P., & Mason, A. P. (2019). Artificial intelligence strategies for national security and safety standards..

Brooke, J. (2020). SUS: A "quick and dirty" usability scale. In *Usability Evaluation In Industry*, pp. 207–212. CRC Press, London, UK.

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, *59*(3), 333–345.

Chatila, R., & Havens, J. C. (2019). *The IEEE global initiative on ethics of autonomous and intelligent systems*, pp. 11–16. Springer International Publishing, Cham.

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165.

Cooke, N. J., & Durso, F. (2007). *Stories of Modern Technology Failures and Cognitive Engineering Successes*. CRC Press.

Coşkun, M., Uçar, A., Yildirim, O., & Demir, Y. (2017). Face recognition based on convolutional neural network..

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests.. *Psychological Bulletin*, *52*(4), 281.

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies*, *XXXII*(1).

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err.. *Journal of Experimental Psychology: General*, *144*(1), 114.

Greene, F., Kudrick, B., & Muse, K. (2014). Human factors engineering at the transportation security administration. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58, p. 2255–2259.

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the 'trust in automated systems survey'? an examination of the jian et al. (2000) scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 217–221.

Hill, K. (2020). Wrongfully accused by an algorithm. In *Ethics of Data and Analytics*, pp. 138–142. Auerbach Publications, New York, NY.

Hill, K., & Mac, R. (2023). Thousands of dollars for something I didn't do. The New York Times. Retrieved from https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Hope, R. M. (2013). *Rmisc: Ryan Miscellaneous.* R package version 1.5.

IEEE SEMVAST Project (2011). IEEE VAST Challenge 2011, Mini Challenge 3 (MC3).. Retrieved May 15, 2023, from https://www.vgtc.org/activities/vastcontest2011/.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, *5*(1), 144–161.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling.* Guilford Publications, New York, NY.

Knop, M., Weber, S., Mueller, M., & Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: Literature review. *JMIR Human Factors*, *9*(1), e28639.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*.

Kriskovic, M., Dutta, S., & Brewer, J. (2017). From dark patterns to angel patterns: Creating trustworthy user experience. *User Experience Magazine*, *16*(5).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lilley, M., Currie, A., Pyper, A., & Attwood, S. (2020). Using the ethical OS toolkit to mitigate the risk of unintended consequences. In Stephanidis, C., Antona, M., & Ntoa, S. (Eds.), *HCI International 2020 – Late Breaking Posters*, pp. 77–82, Cham. Springer International Publishing.

Lim, J., & Cantor, J. R. (2021). Privacy impact assessment for the travel document checker automation using facial recognition. Tech. rep. DHS/TSA/PIA-046(c), PIA-046(c), Department of Homeland Security.

Long, S. K., Sato, T., Millner, N., Loranger, R., Mirabelli, J., Xu, V., & Yamani, Y. (2020). Empirically and theoretically driven scales on automation trust: A multi-level confirmatory factor analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1829–1832.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665.

Microsoft (1995). *The Windows Interface Guidelines—A Guide for Designing Software*. Microsoft Press, Redmond, WA.

ODNI (2015). Intelligence Community Directive 203.. Retrieved from https://www.dni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf.

Palanski, M. E., & Yammarino, F. J. (2011). Impact of behavioral integrity on follower job performance: A three-study examination. *The Leadership Quarterly*, *22*(4), 765–786.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, *50*(3), 511–520.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In Xie, X., Jones, M. W., & Tam, G. K. L. (Eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12. BMVA Press.

Phillips-Wren, G. (2012). AI tools in decision making support systems: A review. *International Journal on Artificial Intelligence Tools*, *21*(02), 1240005.

Qualtrics (2020). *Qualtrics*. Provo, UT. https://www.qualtrics.com.

Raykov, T., & Marcoulides, G. A. (2008). *An Introduction to Applied Multivariate Analysis*. Taylor & Francis Group., New York, NY.

Revelle, W. R. (2018). *psych: Procedures for Personality and Psychological Research*. R package Version 1.18. 10.

Roth, E. M., Bisantz, A. M., Wang, X., Kim, T., & Hettinger, A. Z. (2021). A work-centered approach to system user-evaluation. *Journal of Cognitive Engineering and Decision Making*.

Salehi, P., Chiou, E. K., Mancenido, M., Mosallanezhad, A., Cohen, M. C., & Shah, A. (2021). Decision deferral in a human-AI joint face-matching task: Effects on human performance and trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 65, pp. 638–642. SAGE Publications Sage CA: Los Angeles, CA.

Sargent, R., Walters, B., & Wickens, C. (2023). Meta-analysis qualifying and quantifying the benefits of automation transparency to enhance models of human performance..

SAS Institute Inc. (2023). *JMP®, Version 16*. Cary, NC.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Schaufeli, W. B., Salanova, M., Gonzalez-Roma, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies*, *3*, 71–92.

Schelble, B., Lancaster, C., Duan, W., Mallick, R., McNeese, N., & Lopez, J. (2023). The effect of ai teammate ethicality on trust outcomes and individual performance in human-ai teams..

Schroeder, N. L., Chiou, E. K., & Craig, S. D. (2021). Trust influences perceptions of virtual humans, but not necessarily learning. *Computers & Education*, *160*, 104039.

Sheridan, T. B. (1975). Considerations in modeling the human supervisory controller. *IFAC Proceedings Volumes*, *8*(1, Part 3), 223–228.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. *Data and Policy*, *3*, e3.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52, pp. 1335–1339, Los Angeles, CA. SAGE Publications.

Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. Tech. rep., NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD.

Subirana, I., Sanz, H., & Vila, J. (2014). Building bivariate tables: The comparegroups package for R. *Journal of Statistical Software*, *57*(12), 1–16.

Sung, J., Nguyen, T., Blasch, E., Daniel, C. P. D., G, K., & Mason, A. P. (2019). AI Phase II National Security Standards For Artificial Intelligence: MAST Checklist. Tech. rep., 2019 Public-Private Analytic Exchange Program (AEP).

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using Multivariate Statistics*, Vol. 6. Pearson, Boston, MA.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290.

Wickham, H., François, R., Henry, L., & Muller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8. 1.

Zhu, S., Gilbert, M., Chetty, I., & Siddiqui, F. (2022). The 2021 landscape of FDA-approved artificial intelligence and machine learning-enabled medical devices: An analysis of the characteristics and intended use. *International Journal of Medical Informatics*, *165*, 104828.

Appendix A. The Nine MAST Criteria for Facewise

MAST item	Questions and Feature Descriptions
Sourcing	How well can the system identify underlying sources and methodologies upon which results are based? **High-MAST:** The "View Details" page provides the name of image sources and demographical information about the people whose image data were used to train the AI, such as their race and gender. **Low-MAST:** The system interface does not include the name of image sources and demographical information about the people whose image data were used to train the AI system, such as their race and gender.
Uncertainty	How well can the system indicate and explain the basis for the uncertainties associated with derived results? **High-MAST:** For each pair of images, the system will display a certainty score from 0%-100% to indicate its confidence about its recommended decision. The system also gives an alert if the uncertainty is too high when you click the "Final Decision" button, depending on your decision. Details about how the system calculates the certainty score are available by clicking on the "More Details" button under every decision. The AI's confidence level is calculated in this manner: first, a metric that signifies the mathematical distance between two image pairs is calculated. Then, the difference between the mathematical distance and a pre-determined (computed during the training and validation stages) threshold is calculated. Finally, the difference is normalized by a factor and the confidence level is calculated using probability measures associated with the standard normal distribution. Thus, the AI's confidence is an indication, based on the predetermined threshold. Confidence levels closer to 100% indicate higher confidence. **Low-MAST:** For each pair of images, the system only recommends a binary decision (same or different) and does not indicate its confidence in the decision.
Distinguishing	How well can the system clearly distinguish derived results and underlying data? **High-MAST:** The system can distinguish whether a presented ID is invalid or expired, or if the ID photo may have been digitally altered. An alert message will be automatically shown in these cases by the system. Details about how the system identifies these features in the ID photo are available by hovering over the Crossmark or checkmark icon next to the ID expiration date. **Low-MAST:** The system cannot distinguish whether a presented ID is invalid or expired, or if the ID photo may have been digitally altered.
Analysis of Alternatives	How well can the system identify and assess plausible alternative results?

	High-MAST: In the "View Details" page, the system provides dissimilarity and similarity probabilities as alternatives for each pair. The similarity and dissimilarity numbers are directly derived from the AI's confidence level. The higher of the two probabilities is selected to represent the AI's confidence level. The calculation of the similarity and dissimilarity probabilities assumes that the threshold is distributed as standard normal, and that the scaled differences are realizations of a noise-generating process. Both probabilities are calculated using the scaled difference between the distance metric and the threshold. **Low-MAST:** For each pair of images, the system only gives a decision and does not indicate its confidence in the current decision based on the training and validation stages, nor on probability measures of alternatives associated with the standard normal distribution.
Customer Relevance	How well can the system provide information and insight to users? **High-MAST:** Besides providing the binary decision of same or different, the confidence level, and ID validation on the main page, the system provides additional details through a "More Details" button. This includes information and explanations about similarity, dissimilarity, confidence level, and sources of training for AI. To present the information more efficiently, the system will minimize explanations that have already been shown. Conditional alerts when the system's certainty level is low and alerts about individuals who may need additional screening per the protocol are also included as part of the system with the information displayed as detected. **Low-MAST:** Besides providing the binary decision of the same or different and ID expiration date on the main page, the system does not provide any additional details or any conditional alerts.
Logic	How well can the system help the user understand how it derived its results? **High-MAST:** The system bases its final decision by choosing the larger of similarity and dissimilarity probabilities. "More Details" button also provides an explanation and interpretation of how a prediction or classification is made. Conditional alerts when the system's certainty level is low, and alerts about individuals who may need additional screening per the protocol are also included. To detect the authenticity of an ID photo, a second model was trained, tested, and validated on proprietary datasets of anomalous and non-anomalous travel documents, digitally altered and original images. A separate model further performs character recognition to analyze expiration dates on travel documents. **Low-MAST:** The system does not give any information on how its recommendation is determined. It also does not provide any conditional alerts or any information about the authenticity or validity of the ID photo image.
Change	How well can the system help the user understand how derived results on a topic are consistent with or represent a change from previous analysis of the same or similar topic? **High-MAST:** As you interact with the system, by clicking "more details" you will see a report about your agreement with the system, which indicates how often the system has been uncertain about your final decisions.

	Low-MAST: As you interact with the system, the system does not indicate how often it has been uncertain about your final decisions.
Accuracy	How well can the system make the most accurate judgments and assessments possible, based on the information available and known information gaps? **High-MAST:** For each pair of images, the system will display a certainty score from 0%-100% to indicate its confidence about its recommended decision. System's performance according to the training data and more details about how the system calculates the certainty score are available by clicking on the "More Details" button under every decision. **Low-MAST:** For each pair of images, the system only gives a binary decision and does not indicate its confidence in the decision, the system's performance according to the training data, or more details about how the system made the decision.
Visualization	How well can the system incorporate visual information if it will clarify an analytic message and complement or enhance the presentation of data and analysis? Is visual information clear and pertinent to the product's subject? **High-MAST:** The system automatically shows you an enlarged version of a traveler's ID photo and their photo taken at the security checkpoint. These images will be shown side by side. Distinguishing features that played a big role in determining the recommended decision will also be highlighted by clicking the "View Details" button. **Low-MAST:** The system only shows you an enlarged version of a traveler's ID photo and their photo taken at the security checkpoint without any additional visualized explanation about the recommended decision.

Table 2: MAST (Blasch et al., 2021) and Facewise feature descriptions for High-MAST and Low-MAST.

Appendix B: The Nine MAST Criteria for READIT

MAST item	Questions and Feature Descriptions
Sourcing	How well can the system identify underlying sources and methodologies upon which results are based? **High-MAST:** In the documents page, you can see descriptive information about the documents (data) used to gather the clusters including basic information and detailed descriptions of the sources. The datasheet for READIT includes information on the clustering model, models for summarization, training data, possible biases, preprocessing of data, and quality of the data used in training to derive results. In the main dashboard view, you can view the data used to derive the cluster either by hovering or clicking on it including the cluster title, number of documents, top terms, and representative documents. The representative documents can be viewed as a summary (derived result) or raw version. **Low-MAST:** For any given cluster in the main dashboard view, you can view more details about it by clicking on it. The title of the cluster, number of documents, and summaries of the documents will be displayed in the documents and summaries pane. Only the derived results are shown, not the underlying sources and data used to derive the clusters or summaries.
Uncertainty	How well can the system indicate and explain the basis for the uncertainties associated with derived results? **High-MAST:** READIT indicates levels of uncertainty with derived results in two ways, as described in the datasheet. First, READIT includes keywords per cluster to show how documents in clusters are related to each other. Keywords are displayed with a term frequency–inverse document frequency (tf-idf) score which measures the certainty the word fits with the cluster. Second, READIT includes similarity scores to assess the similarity between clusters. This score is calculated using cosine similarity to show the certainty that clusters are related to each other. **Low-MAST:** In the topic similarity visualization, the relationship between two topics is colored from white to dark blue with dark blue indicating a higher certainty the two topics are related. These relationships are not labeled with numbers, neither is it explained how this similarity is calculated.
Distinguishing	How well can the system clearly distinguish derived results and underlying data? **High-MAST:** For any given cluster you can view more details about the data used to derive the cluster either by hovering or clicking on it. The datasheet for READIT includes information on the clustering model, models for summarization, training data, underlying assumptions for choice of training data, quality of the data used in training to derive results, possible biases, pre-processing of data, recommended uses and users, and restrictions on use. The datasheet was created with domain expert input.

	Low-MAST: When opening or clicking on clusters, you can view more details about that cluster. The title and summary of representative documents will appear. The raw data used to derive the title and summaries is not displayed. There is no datasheet with information on how these titles or summaries are calculated.
Analysis of Alternatives	How well can the system identify and assess plausible alternative results? **High-MAST:** In the topic similarity visualization, users initially view the visualization where the topics are ordered alphabetically. By factoring in the similarity score and uncertainties, READIT can reorder the view in this visualization such that highly related topics will appear together to present an alternative view. **Low-MAST:** READIT is not able to show alternative results when uncertainties in the data warrant them. There is no way to reorder visualizations based on any criteria.
Customer Relevance	How well can the system provide information and insight to users? **High-MAST:** READIT synthesizes large corpora of documents and produces clusters of similar documents. The topic similarity visualization shows which clusters are most highly related to each other. Users can examine the clusters and their relationships in the topic similarity view for trends for follow-up work. READIT is also able to suggest locations to filter by if the documents contain multiple locations. Users can also filter all visualizations by topic. There is a topic filtering pane where users can check all, or some topics and the corresponding selected topics will be highlighted in the visualizations. **Low-MAST:** READIT synthesizes large corpora of documents and produces clusters of similar documents. The topic similarity visualization shows which clusters are most highly related to each other. Users can examine the clusters and their relationships in the topic similarity view for trends for follow-up work.
Logic	How well can the system help the user understand how it derived its results? **High-MAST:** For any given cluster you can view more details about the data used to derive the cluster either by hovering or clicking on it. The datasheet includes information on pre-processing of data. READIT includes an option to filter results by location, if location information is detected in the document. To give location options, READIT must consider the location information in the context of the document, and other assumptions about the embedding of the location in the document. **Low-MAST:** When clicking on clusters in the main view, you can view the title and representative documents in summary form. The titles and summaries are understandable to users. Information on how clusters, titles, and summaries are formed is not included. There is also no information on the pre-processing of data.
Change	How well can the system help the user understand how derived results on a topic are consistent with or represent a change from previous analysis of the same or similar topic?

	High-MAST: In the documents page, READIT includes information on similar searches from other agencies. Similar searches may be based on the average length of the document, number of documents, or number of clusters generated. **Low-MAST:** READIT does not have a way to note changes from previous analyses or similar analyses. READIT also cannot compare current results with those of other agencies which had similar results.
Accuracy	How well can the system make the most accurate judgments and assessments possible, based on the information available and known information gaps? **High-MAST:** The READIT datasheet includes information on system verification and validation methodology, and results from the training data where the system achieved sufficiently high accuracy. To assess the accuracy of READIT, users can view the full documents used in each cluster and compare them against the top terms to independently determine whether the documents match the top terms. Likewise, users can view a summary of the document and compare it against the full version of the document in the documents and summaries view to see if the summary is accurate. **Low-MAST:** READIT does not include information on system verification, validation methodology, or information on the training of the system where it achieved sufficient accuracy. Since underlying sourcing information and raw data are not included in the system, it is difficult to assess whether the topics and summaries are accurate.
Visualization	How well can the system incorporate visual information if it will clarify an analytic message and complement or enhance the presentation of data and analysis? Is visual information clear and pertinent to the product's subject? **High-MAST:** READIT uses three main visualizations to enhance users' understanding of the clusters. First, in the topic overview visualization, clusters are displayed as bubbles where the size of the bubbles can indicate anomalies. Next, READIT also creates and displays a topic similarity visualization to help understand the connections between clusters. Lastly, there is a timeline view in READIT to display clusters on a timeline (if documents contain date information). All visualizations are simple and labeled properly. Users can view more details about the visualizations by clicking on them or hovering over them or filtering all visualizations by cluster using the filtering option. **Low-MAST:** READIT uses two visualizations. The similarity matrix shows the similarity scores between topics. Darker colors indicate more similarity but score values are not shown. The timeline shows the clusters on the timeline. Visualizations contain no interactivity and users are not able to click or hover on items to view more details about the visualizations.

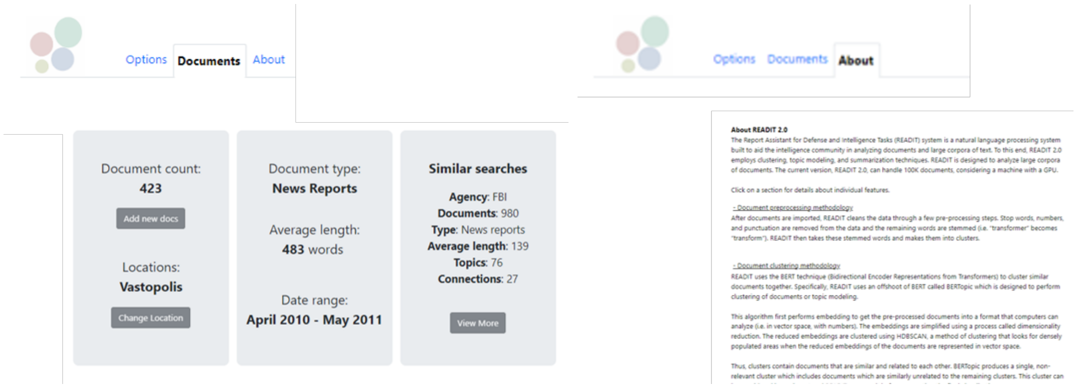
Table 3: MAST (Blasch et al., 2021) and READIT feature descriptions for High-MAST and Low-MAST.

Appendix C: Study Questionnaires

Variables	**Reference**	**Example item(s)**	**Number of items/ Reverse items**	**Scale**	**Facewise Cronbach's Alpha**	**READIT Cronbach's Alpha**
MAST-total	(Blasch et al., 2020)	Sourcing, uncertainty, distinguishing, analysis of alternatives, customer relevance, logical argumentation, consistency, accuracy, and visualization	9/0	9 - 36	.91	.91
Risk	(Weber et al., 2002)	Please indicate how risky you perceive it is to use this system for completing your task well.	1/0	1 - 5	-	-
Benefit	(Weber et al., 2002)	Please indicate how beneficial you perceive it is to use this system for completing your task well.	1/0	1 - 5	-	-
Trust (Jian)	(Jian et al., 2000)	"I can trust the system."; "The system looks deceptive."	12/5	1 - 7	0.90	0.92
Trust (Chancey)	(Chancey et al., 2017)	"I understand how the system will help me perform well. "; "The information the system provides reliably helps me perform well.	15/0	1 - 7	0.96	0.96
Credibility	(Appelman & Sundar, 2016)	"How accurate do the results of the system appear to be?"; "How believable do the results of the system appear to be?"	3/0	1 - 7	0.92	0.92
Engagement	(Schaufeli et al., 2002)	"I was immersed in this research task."; "To me, this research task was challenging."	17/0	1 - 7	0.91	0.93
Usability (SUS)	(Brooke, 2020)	"I felt very confident using the system."; "I thought the system was easy to use."	10/5	1 - 5	0.80	0.88
Task performance	-	Average response time and Accuracy for Facewise and final report gradings for READIT	2/0	0	-	-

Table 4: Dependent and Control Variables.

Appendix D: READIT Documents and About Tabs for the High-MAST Version



Document count:
423

Locations:
Vastopolis

Document type:
News Reports

Average length:
483 words

Date range:
April 2010 - May 2011

Similar searches

Agency: FBI
Documents: 980
Type: News reports
Average length: 139
Topics: 76
Connections: 27

About READIT 2.0
The Report Assistant for Defense and Intelligence Tasks (READIT) system is a natural language processing system built to aid the intelligence community in analyzing documents and large corpora of text. To this end, READIT 2.0 employs clustering, topic modeling, and summarization techniques. READIT is designed to analyze large corpora of documents. The current version, READIT 2.0, can handle 100K documents, considering a machine with a GPU.

Click on a section for details about individual features.

- Document preprocessing methodology

After documents are imported, READIT cleans the data through a few pre-processing steps. Stop words, numbers, and punctuation are removed from the data and the remaining words are stemmed (i.e. "transformer" becomes "transform"). READIT then takes these stemmed words and makes them into clusters.

- Document clustering methodology

READIT uses the BERT technique (Bidirectional Encoder Representations from Transformers) to cluster similar documents together. Specifically, READIT uses an offshoot of BERT called BERTopic which is designed to perform clustering of documents or topic modeling.

This algorithm first performs embedding to get the pre-processed documents into a format that computers can analyze (i.e. in vector space, with numbers). The embeddings are simplified using a process called dimensionality reduction. The reduced embeddings are clustered using HDBSCAN, a method of clustering that looks for densely populated areas when the reduced embeddings of the documents are represented in vector space.

Thus, clusters contain documents that are similar and related to each other. BERTopic produces a single, non-relevant cluster which includes documents which are similarly unrelated to the remaining clusters. This cluster can

(a) Documents tab

(b) About tab

Figure 9: Documents and About tabs in High-MAST READIT platform.

Appendix E: 95% Confidence Interval Figures

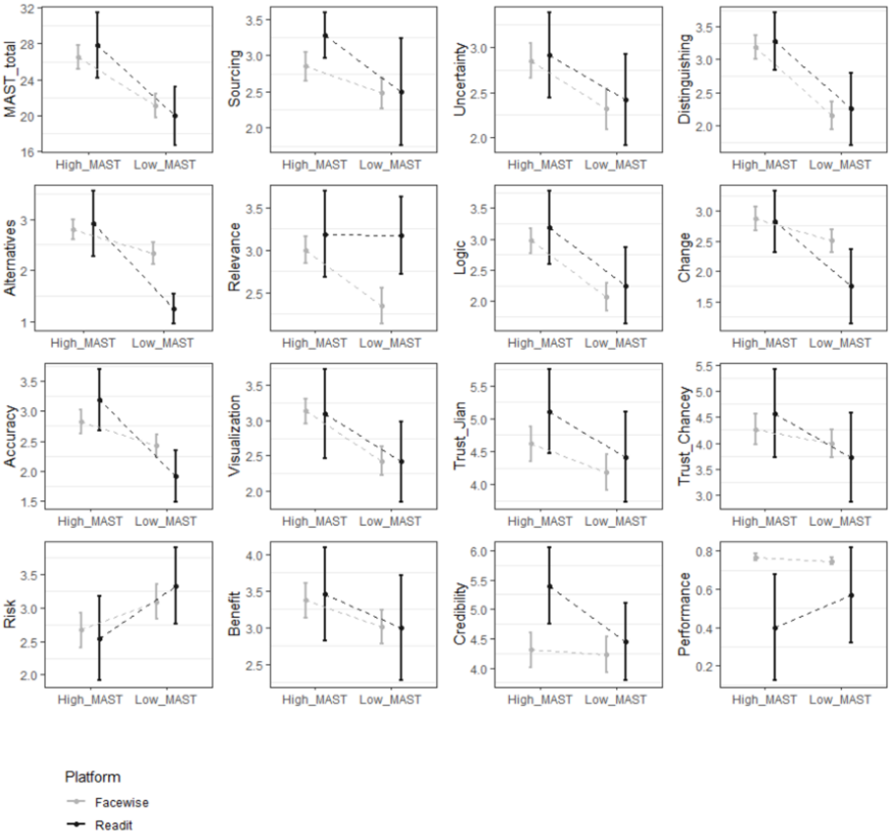


Figure 10: Means with 95% Confidence Intervals for Facewise and READIT across different levels of Low-MAST and High-MAST. We used -●- Facewise for Facewise and -●- Readit for READIT.

Appendix F: Participant demographics for Facewise and READIT

	High-MAST ($n = 73$)	**Low-MAST** ($n = 73$)
Years of experience as a TSO	55% 3 years or less 24% 10 or more years	50% 3 years or less 28% 10 or more years
Highest degree	69% 2-year college or less 26% 4-year college	74% 2-year college or less 21% 4-year college
Volunteer hours in the past 3 months	62% 0 hours	71% 0 hours
Computer habit	58% daily	65% daily
Gaming habit	18% daily 26% never 26% never	30% daily 17% never
Screen hours before study	Mean: 2 hrs. Median: 1.2 hrs.	Mean: 2.2 hrs. Median: 2 hrs.

Table 5: Participant demographics across High-MAST and Low-MAST for Facewise.

	High-MAST ($n = 11$)	**Low-MAST** ($n = 12$)
Age	36% 30 years or less 18% 31-39 years 46% 40 or more years	34% 30 years or less 33% 31-39 years 33% 40 or more years
Gender	73% man 27% woman	50% man 50% woman
Race	82% white	83% white
Years of experience as an IA	18% 2 years or less 27% 3-5 years 55% 6 years or more	25% 2 years or less 17% 3-5 years 58% 6 years or more
Experience with AI-DSS	46% no prior experience	33% no prior experience
Highest degree	27% 4-year college 73% master's	17% 4-year college 66% master's 17% doctorate
Experience with VAST challenge	100% no	100% no
Experience with clustering tools	55% no	33% no
Screen hours before study	Mean: 5.5 hrs. Median: 6 hrs.	Mean: 5 hrs. Median: 5 hrs.

Table 6: Participant demographics across High-MAST and Low-MAST for READIT.