



Trust Factors in Voice User Interfaces for International Nuclear Safeguards Inspectors



Presented by: Jamie Coram

With: Kristin Divis, Breannan Howell, and Zoe Gastelum

July 2024

SNL Human Systems Symposium



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

International Nuclear Safeguards are voluntary measures for states implemented by the International Atomic Energy Agency (IAEA, under auspices of the United Nations) to **detect** and **deter** misuse of civilian nuclear programs.

States declare, inspectors verify

Information from a state (e.g., inventory, plans)



Information collected in the field (e.g., measurements, samples, images)



Other information analyzed at the IAEA (e.g., open-source satellite imagery, trade documents)



(Photos: IAEA)



Motivation

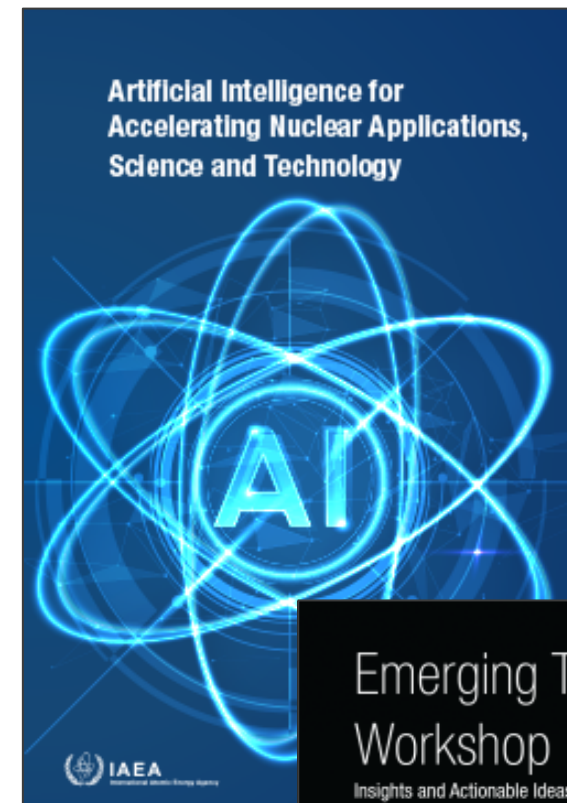
Safeguards workload is increasing, and Artificial Intelligence (AI) has the potential to help.

Sandia has previously studied use of *visual* interfaces with AI, but **voice user interfaces (VUIs)** are new for safeguards.

If inspectors are going to **use** VUIs, then they need to **trust** them.

Nuanced challenges justify domain-specific experimental validation:

- High consequence
- Time pressure
- Budget-constrained
- Divided attention
- Jetlag
- Non-native language
- Use of personal protective equipment
- Loud/industrial environments
- Potential hazards



Trust in Voice User Interfaces (VUIs)

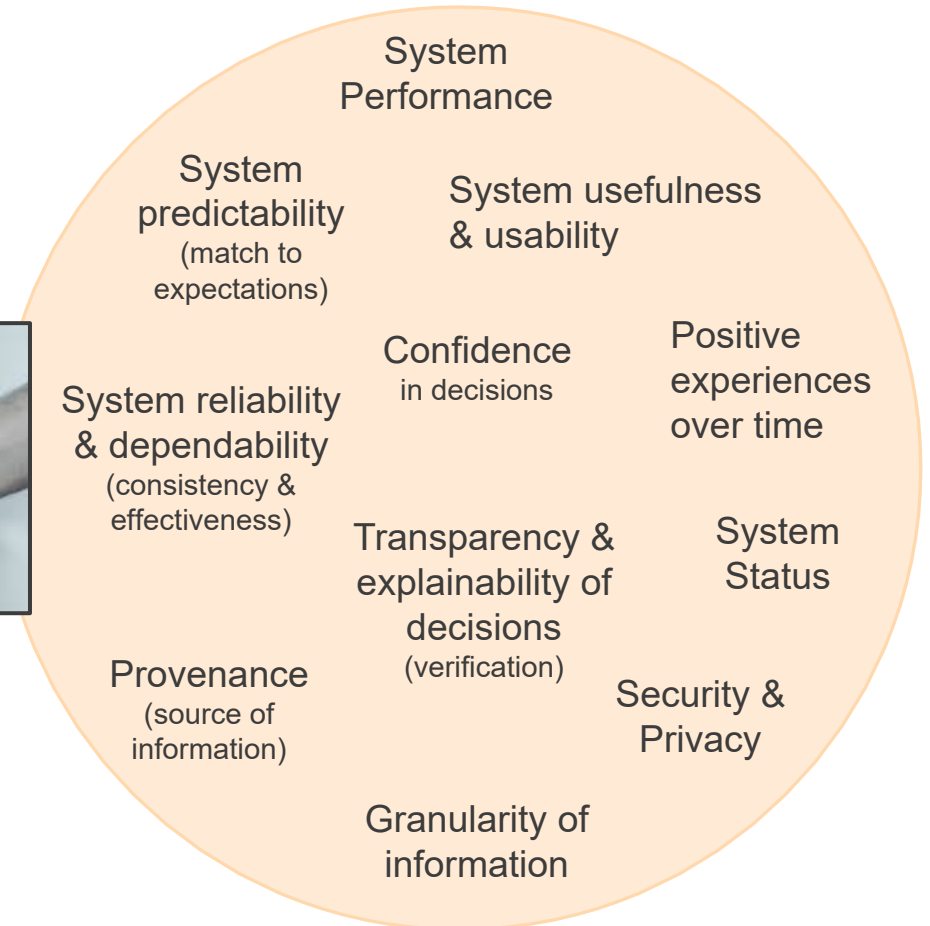


Tasks VUIs might perform:

- Read information
- Communicate decisions or analysis results from underlying models
- Request clarification or additional information
- Provide step-by-step directions or task tracking
- Confirm input or receipt of information
- Record user observations
- Facilitate communication among team members
- Provide alerts or notifications
- Refer user to another platform
- Pause, exit, or switch tasks
- Provide application status



Factors that Impact Trust:



Goal: Provide *actionable recommendations* to VUI developers to “right size” trust

Task 1:

Seal Examination – Inspectors **secure containers with seals** that indicate if containers have been **opened** since the last inspection.



(Photos: IAEA)

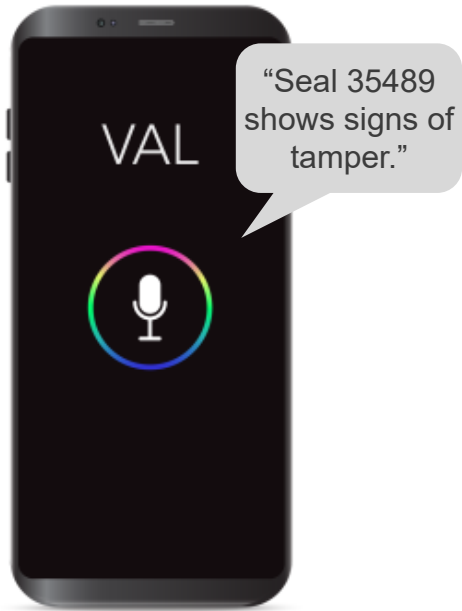
Task 2:

Material Measurement – Inspectors use detectors to **measure the amount of nuclear materials** in containers.

Seal Examination Task*



Participants decide whether to **keep** or **remove & replace** seals, with assistance from VAL, our simulated voice user interface



Simulated Voice
Assistant Laboratory
(VAL)



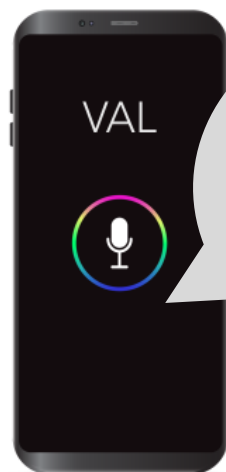
Example: Normal wear &
tear (keep)



Example: Tamper scratches
(remove & replace)

* Please note that the tasks and stimuli used in this experiment are simplified versions intended for testing purposes only. They do not necessarily reflect real indications of nuclear misuse. Study controlled for normal wear & tear and tamper patterns, including location and difficulty (independently normed).

Location Explainability



“Seal 46184 shows signs of tamper on the front lower left”

Granularity of location information

- *None*: No location information
- *Medium*: e.g., “signs of tamper on the front”
- *High*: e.g., “signs of tamper on the front lower left”

Confidence



“Seal 46184 shows signs of tamper with 70% confidence”

Granularity of confidence information

- *None*: No confidence information
- *Low*: Text only (e.g., “medium confidence”)
- *Medium*: Numerical increments of 5% (e.g., 65%)
- *High*: Numerical increments of 0.1% (e.g., 64.9%)



VAL did not always provide the correct information.

	VAL indicates “signs of tamper”	Val indicates “normal wear & tear”	
Seal has tamper	True positive (n=20) <i>Remove the seal</i>	False negative (n=5) <i>Remove the seal</i>	
Seal does not have tamper	False positive (n=10) <i>Keep seal in place</i>	True negative correct ID (n=32) <i>Keep seal in place</i>	True negative incorrect ID (n=8) <i>Remove the seal</i>

75 experimental trials per participant

VAL’s accuracy on the **tamper detection task** was **80%**

Sometimes VAL provided an **incorrect seal ID**. VAL’s accuracy on **both tasks** was **69%** (80% for just tamper detection; 89% for just ID)

Location or confidence information was only provided *when VAL indicated signs of tamper*



Data collected using Pavlovica with Prolific*

- Participants in the US & fluent in English with a 90% minimum prior approval rate
- \$6 for 15-30 minutes of participation, plus \$0.03 bonus for each correct trial (up to \$2.25)

Behavioral performance metrics on **each trial decision**

- **Accuracy** on the task (tamper detection and correctness of seal ID)
- **Compliance** with VAL's analysis – indicative of trust, especially over-trust when VAL is incorrect
- **Response time**, as a surrogate for search or deliberation time – also indicative of trust

Subjective ratings of trust in the VUI **self-reported via post-task questionnaire**

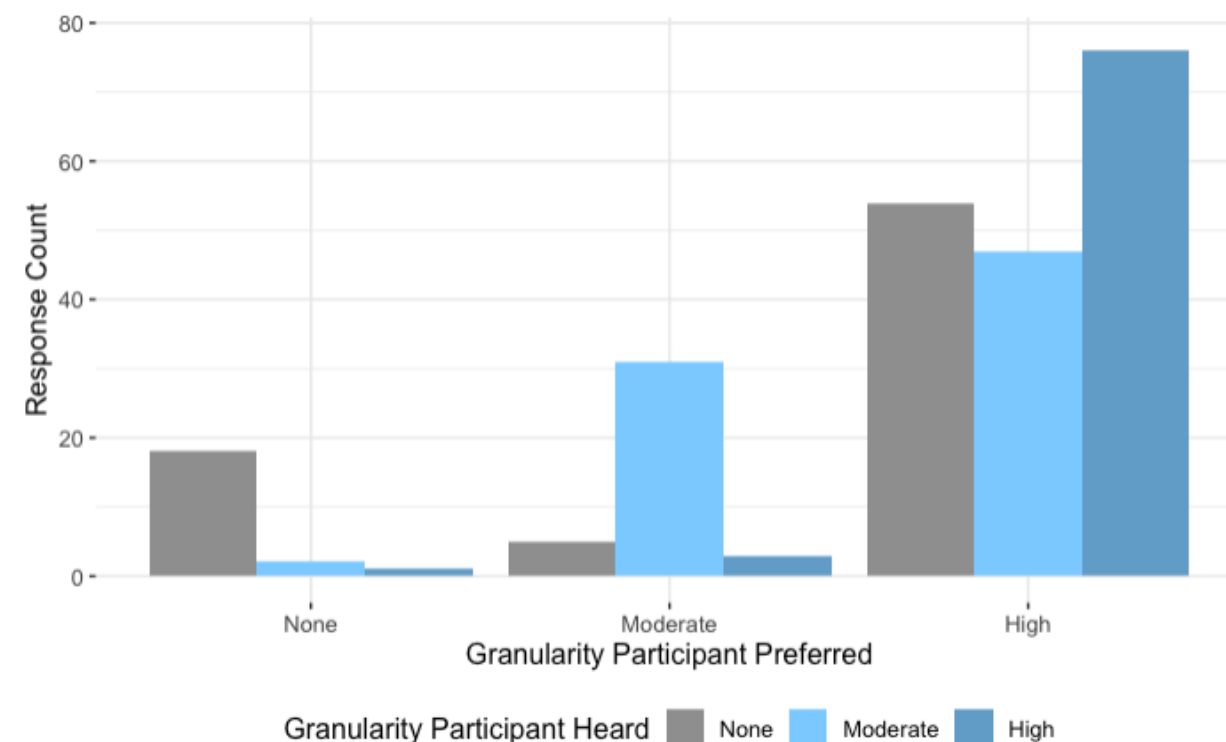
- Project-specific – e.g., overall trust and reliability ratings, influence of and preference for trust factor manipulations
- From literature – e.g., TOAST (Trust of Automated Systems Test) and ATI (Affinity for Technology Interaction)

** Location Explainability: 237 participants across 6 between-subjects lists. Confidence: 158 participants across 8 between-subjects lists. Participants in both studies were 50% male, 50% female, mean age ~37 with std dev ~12 years (of those who reported their gender and age)*

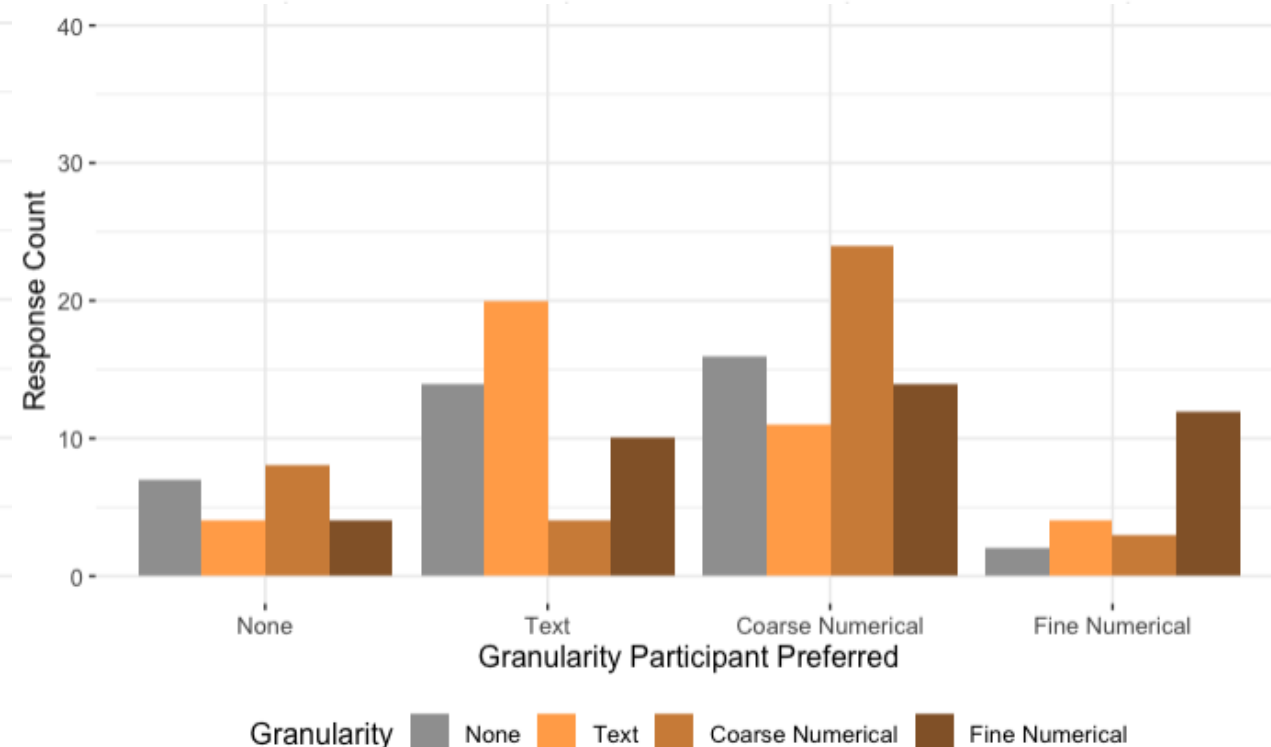
Granularity Preference



Location Explainability



Confidence

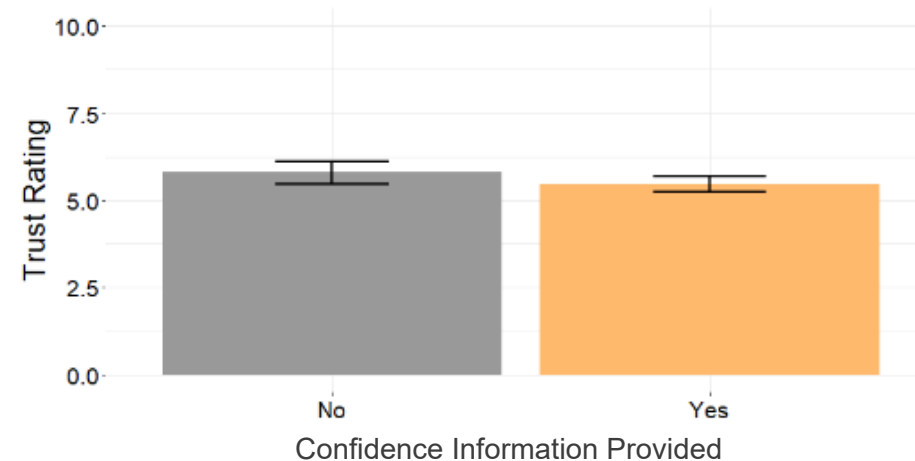
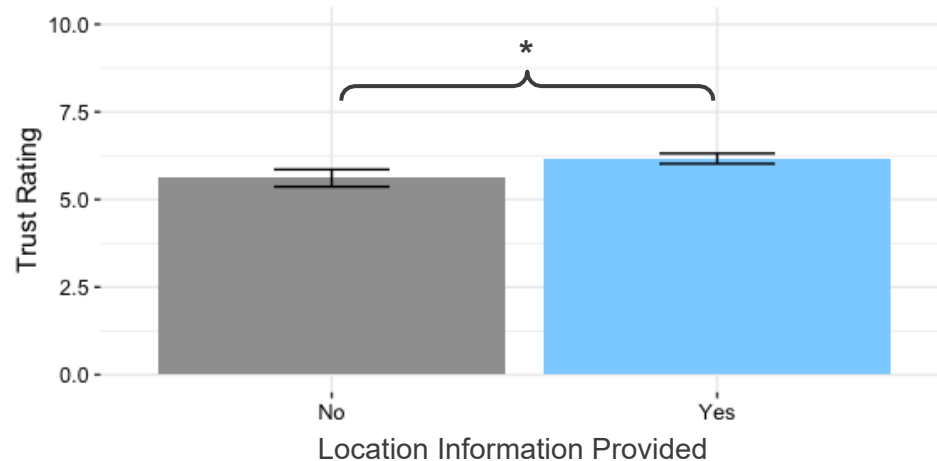


- Most participants preferred **high granularity** location information and either **text** or **coarse numerical** confidence information
 - In both studies, there was some preference for the variant the participant heard

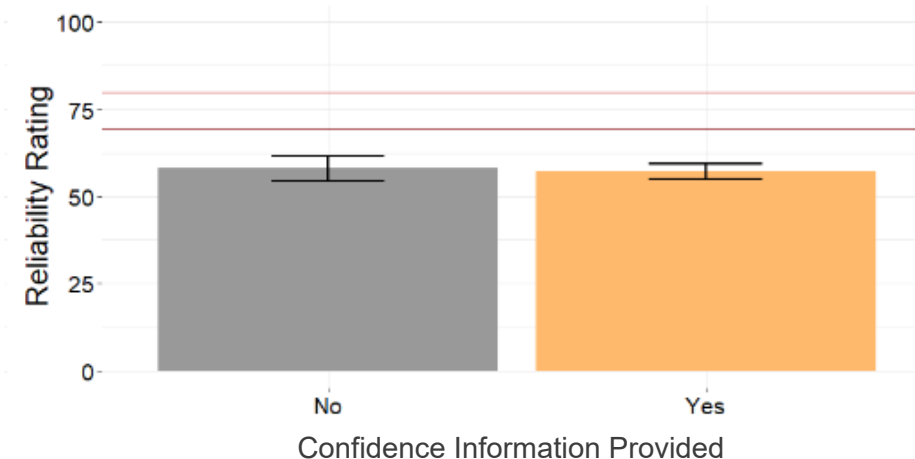
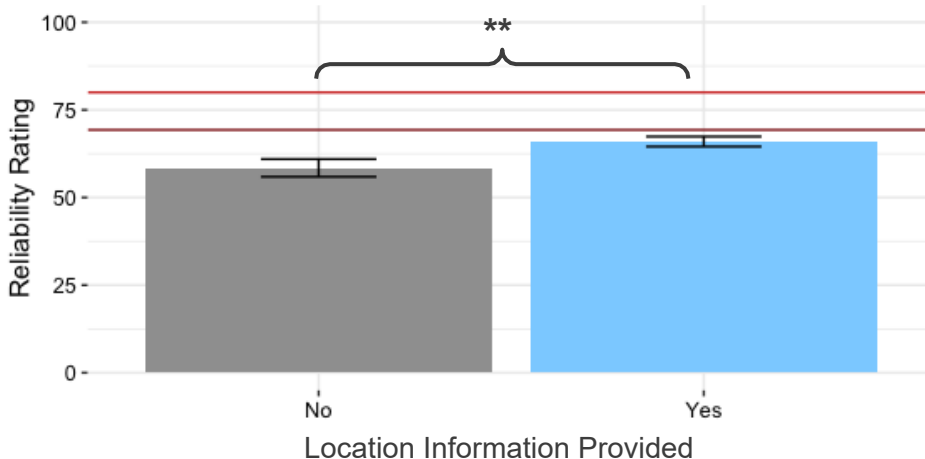
Subjective Trust & Reliability Ratings



Trust



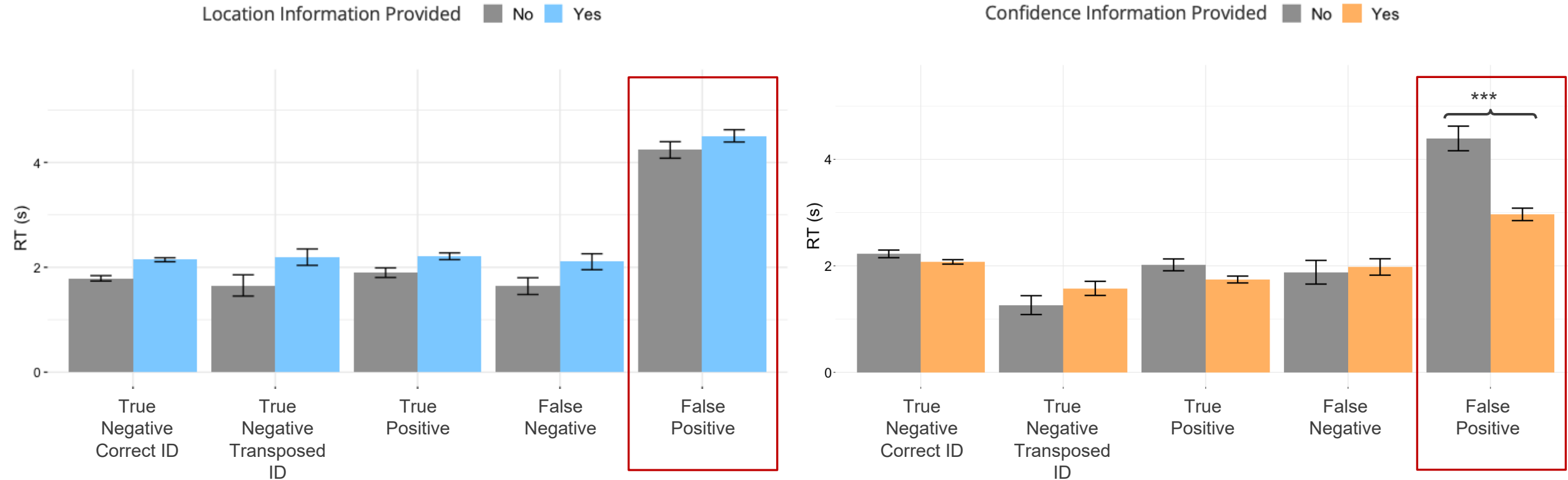
Reliability



* = $p < .05$
** = $p < .01$
*** = $p < .001$

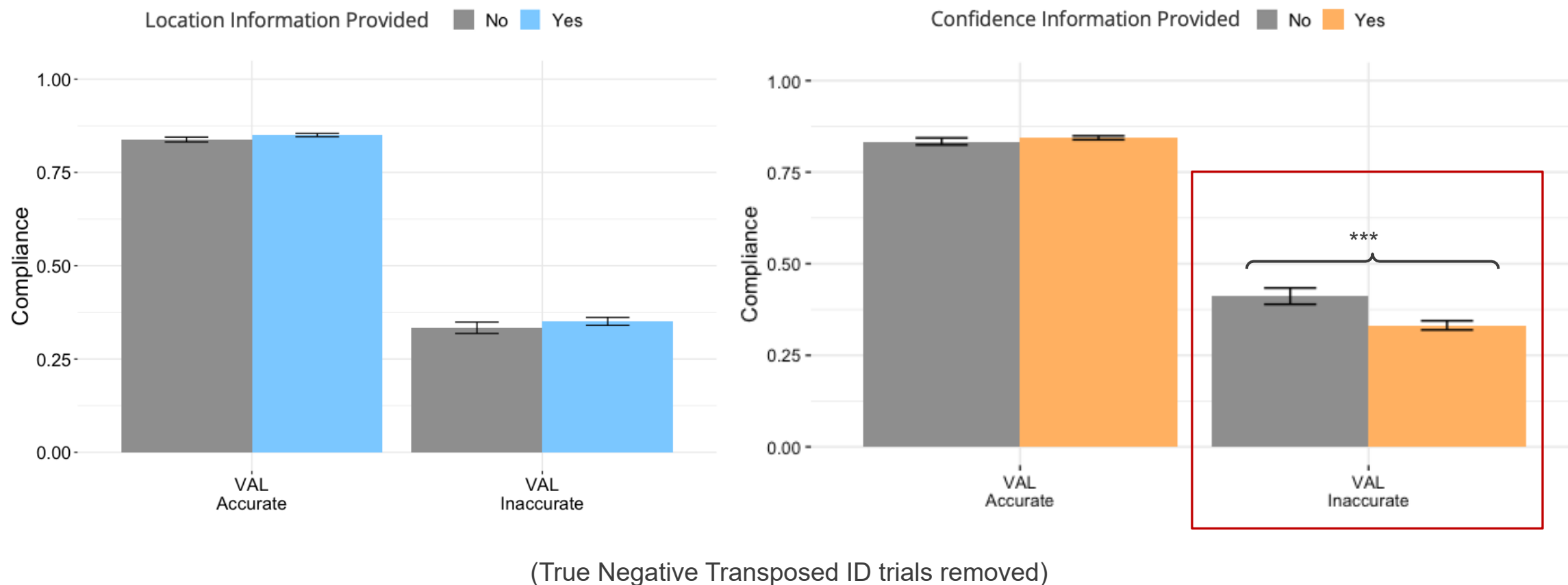
- Participants rated trust and reliability similarly and tended to **underestimate VAL's reliability**
- Participants who received **location** information reported slightly higher **trust** and **reliability** ratings

Response Time on Accurate Trials



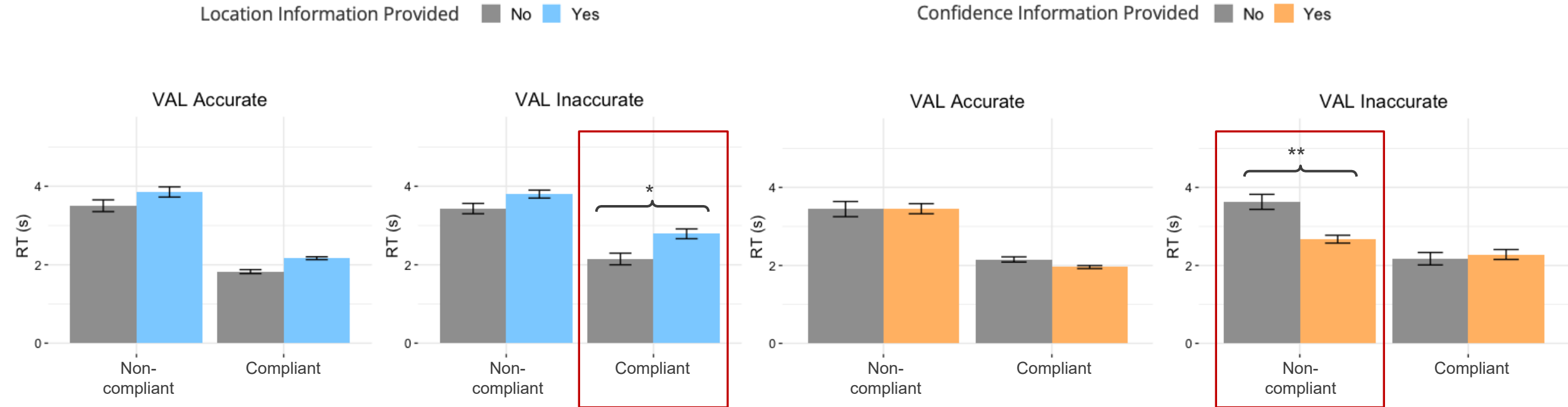
- Participants spent more time looking for a tamper when VAL falsely indicated one was present
 - Participants who received **confidence** information responded **faster** to false positive seals

Compliance Rate



- Participants complied with VAL in about one third of trials where VAL made an error
- Participants who received **confidence** information were less likely to comply when VAL made an error

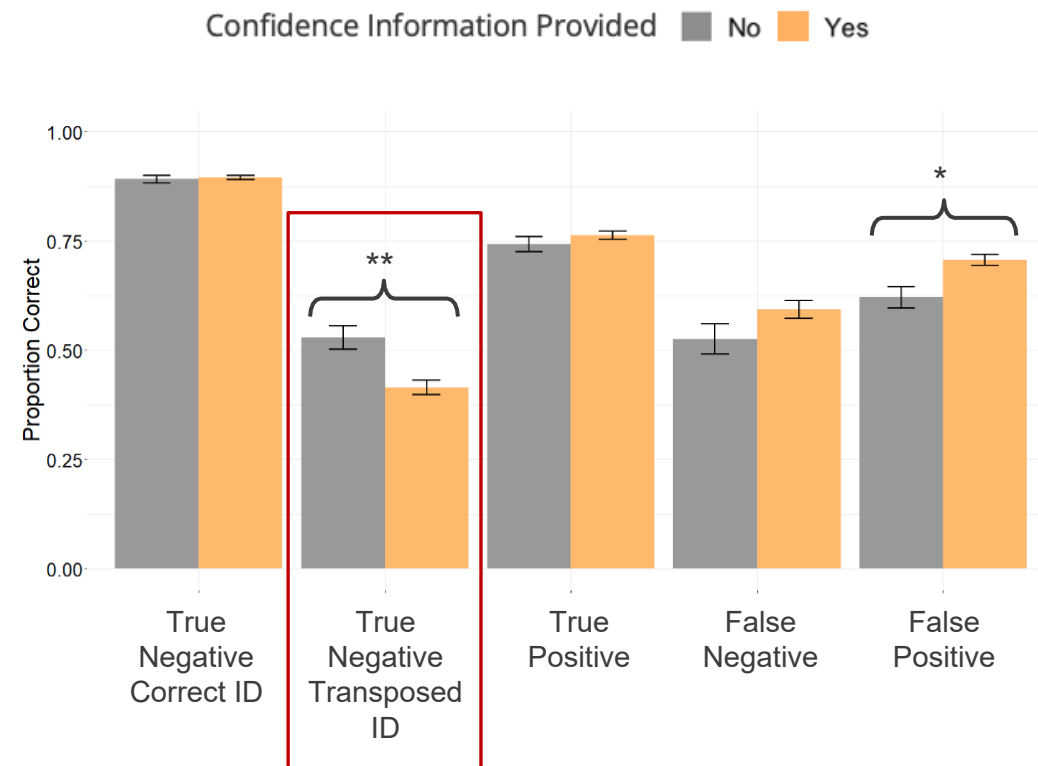
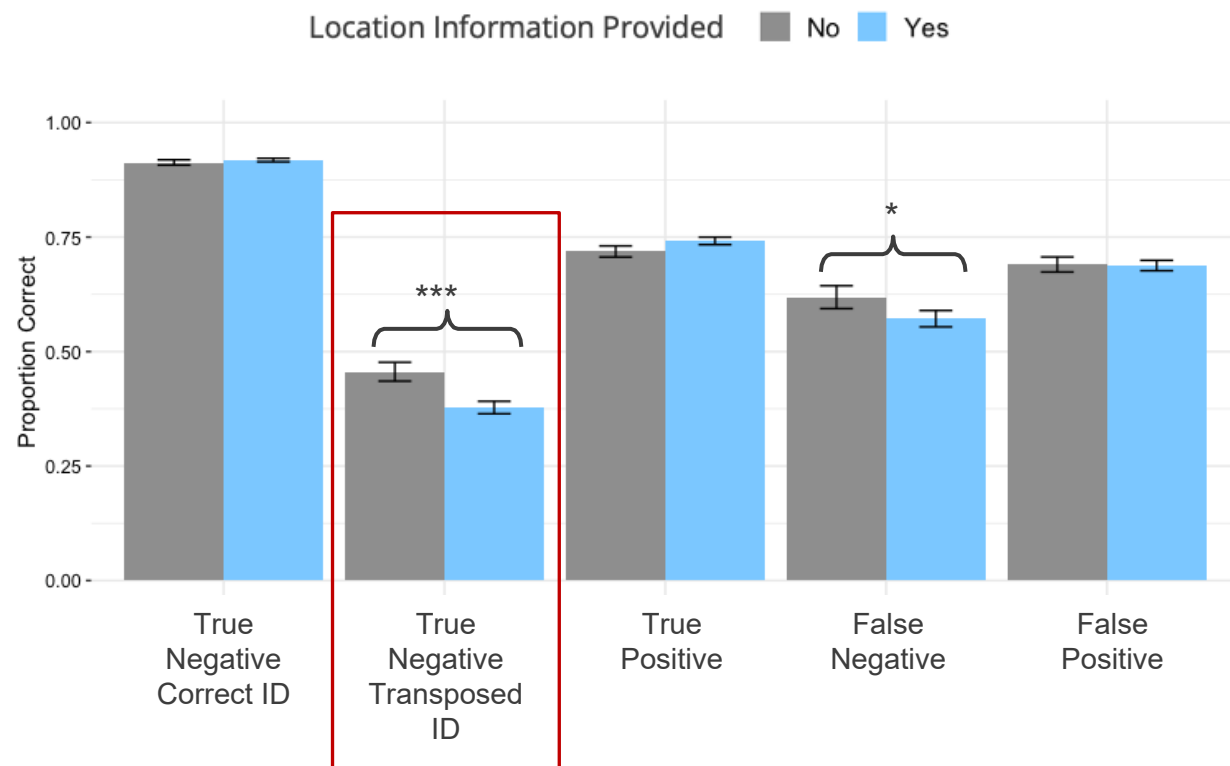
Response Time vs. Compliance



(True Negative Transposed ID trials removed)

- Participants who received **location** information were **slower** to respond when incorrectly complying with VAL
- Participants who received **confidence** information responded **faster** when correctly disagreeing with VAL





Incorrect Seal IDs



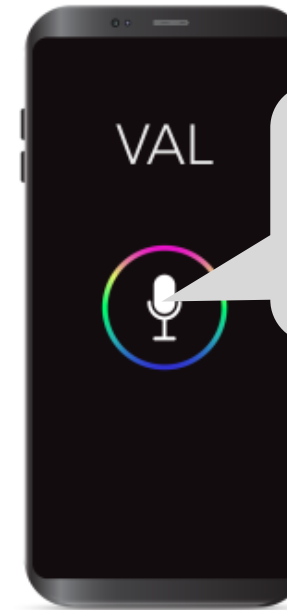
(Keep in mind VAL only provided location or confidence information on **True Positive** and **False Positive** trials)

- Participants often missed incorrect seal IDs
- Participants who received **location** or **confidence** information were *more* likely to miss incorrect seal IDs



Finding	Recommendation
<p>Participants were slower to respond to false positive seals and when disagreeing with VAL – possible indicators of trust.</p> 	<p>Provide confidence information to improve accuracy and response time and to help users more appropriately calibrate trust <i>when the VUI is incorrect on the primary task</i> (false positives).</p>
<p>Participants often missed incorrect seal IDs – with accuracy further reduced when providing location or confidence information.</p> 	<p>Insulate secondary tasks from performance reduction in situations where the VUI strongly emphasizes the primary task.</p>
<p>Participants reported fairly low trust ratings and underestimated VAL's reliability.</p> 	<p>Provide explainability information (e.g., location) to help calibrate users' overall trust and reliability perceptions of the VUI.</p>
<p>Manipulating the granularity of location or confidence information did not significantly impact the effects in these studies.</p> 	<p>Provide levels of granularity appropriate to the task and consider participants' preference for moderate to high granularity information.</p>

- Currently analyzing data from *Material Measurement* studies
 - Simulating the use of detectors to **measure the amount of nuclear materials** in containers
 - Manipulating **granularity** of measurement and **provenance** information (prior measurements)
- Proposed follow-on work to study data provision across **multiple data modalities** (e.g., visual, audio, haptic, robotic interfaces...)



“Container 18150 measures 3.1% enriched. This container measured 3.0% enriched during a prior inspection.”





Thank you! Questions?

Jamie Coram | jcoram@sandia.gov

