RELIABILITY REQUIREMENTS FOR AUGMENTED REALITY IN VISUAL SEARCH TASKS

Dr. Sam Monfort, Dr. John Graybeal

CCDC C5ISR Center Night Vision and Electronic Sensors Directorate

24 October 2019

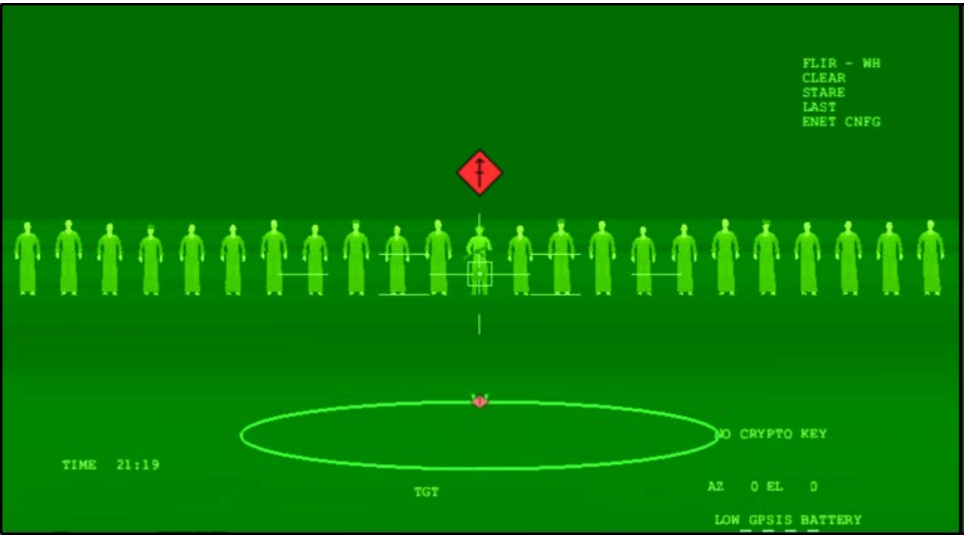
AUGMENTED REALITY

- **Augmented reality (AR) technologies have great potential to improve battlefield performance**
- **Soldiers must process information from an outside source and integrate it into their decision making**
- **AR that fails to provide correct information, or provides incorrect information, may harm performance**
  - e.g., unnoticed failures, distractions, distrust of accurate information, etc.

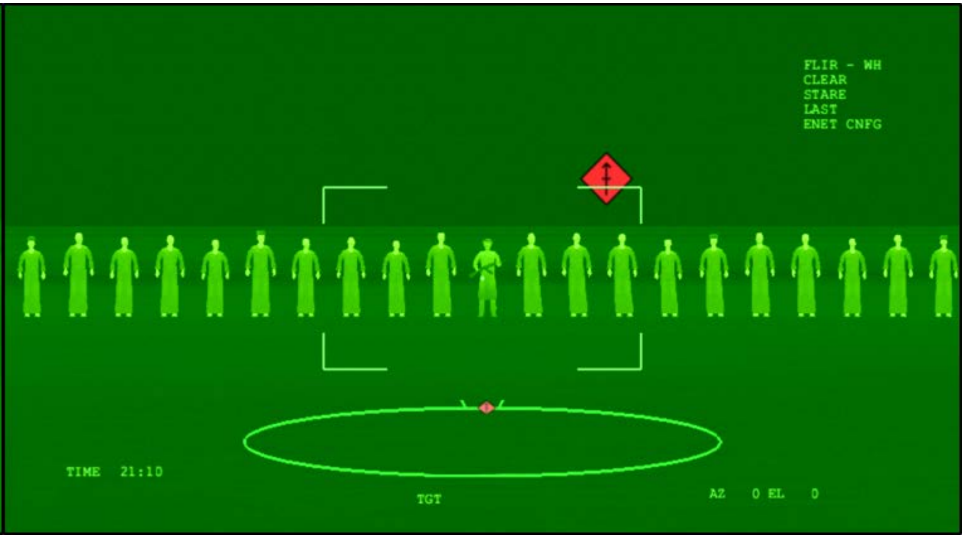
INTRODUCTION

- **Augmented reality-aided target ID will not be perfect** **(Biros, Daly, & Gunsch, 2004)**
- **Soldier trust is required for use/adoption of new technology, as distrust = disuse** **(Parasuraman & Riley, 1997)**
- **What level of AR accuracy is necessary?**
  - …to improve human performance?
  - …to facilitate trust?

ACCURATE



INACCURATE



HYPOTHESES

- **Errors in the false-alarm prone AR will be more damaging to both objective performance and subjective state than miss-prone AR**

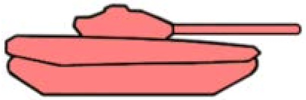
FALSE ALARM

MISS

- **Errors (either type) above distant targets will be more damaging to objective performance than errors above close targets**

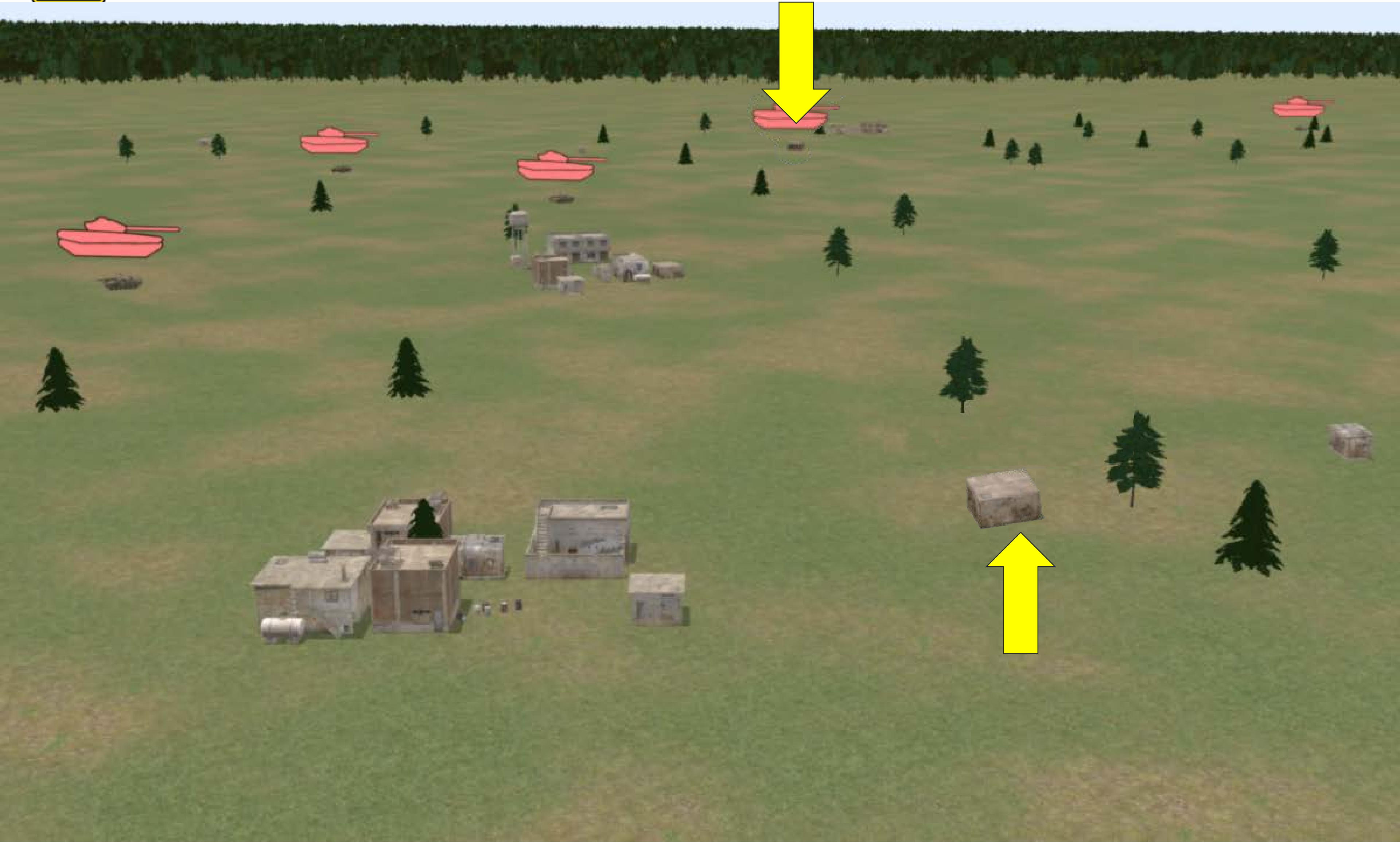
METHOD

- **Participants asked to spot tanks in 54 consecutive grassland scenes**
  - Each scene contained between 0 and 8 targets
- **Search task guided by intuitive AR icons:**



- **Participants assigned to one AR error-type condition: false-alarm prone or miss-prone**
- **AR reliability varied throughout: {25%, 40%, 55%, 70%, 85%, 100%}**
  - Reliability corresponded to number of AR mistakes in a scene

MISS AND FALSE ALARM EXAMPLES



PARTICIPANTS

- **We used a simple visual perception task that did not require previous experience (so anyone could participate)**
  - Sample should match population on relevant variables

- **Total of 184 participants recruited in person and over the Internet**

- **Internet participants excluded for poor screen resolution (n=32) or for not finishing the task (n=12)**

Final Sample
n=140

n=83
Given course credit

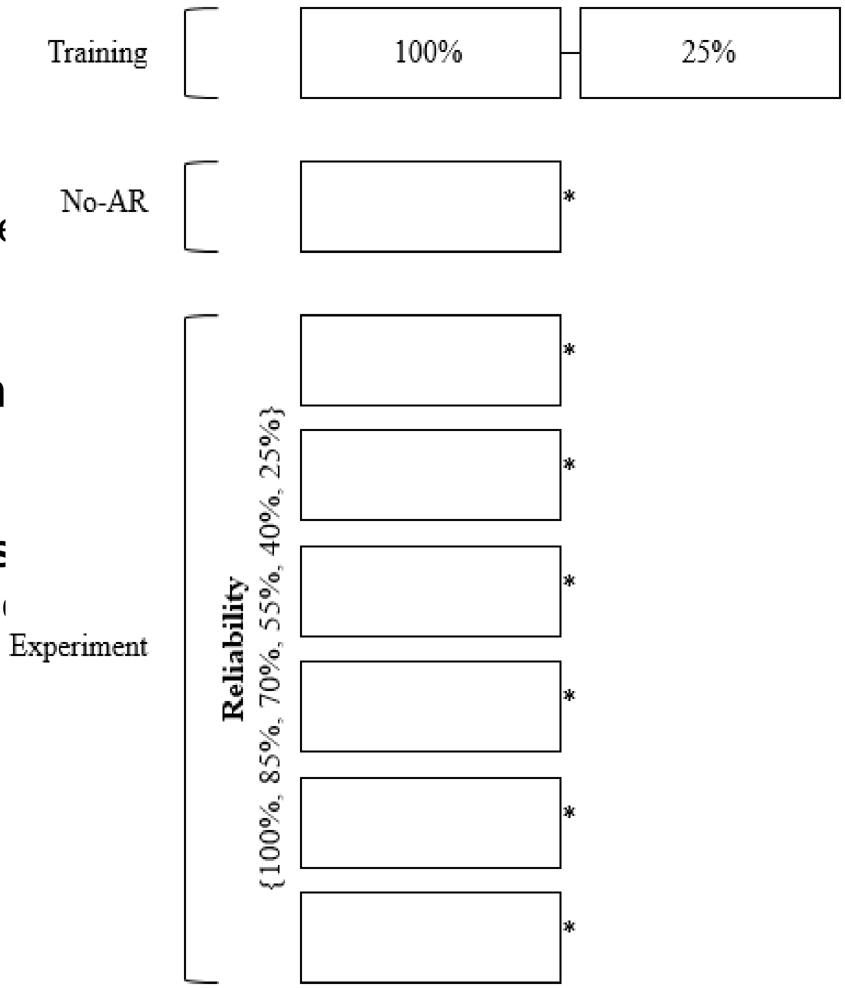
n=57
Paid $3

SELF-REPORT SURVEYS

- **Useful for describing how participants experience the task**
- **Similarities/differences between objective and subjective metrics are informative (i.e., not recognizing safety hazards)**
- **Three self-report measures:**
  - Survey on Trust in AR *(“How much do you trust the AR to help you?”)*
  - Overall workload scale from NASA Task Load Index *(“How hard was that?”)*
  - Gas Tank Questionnaire *(“How much energy do you have left?”)*

PROCEDURE

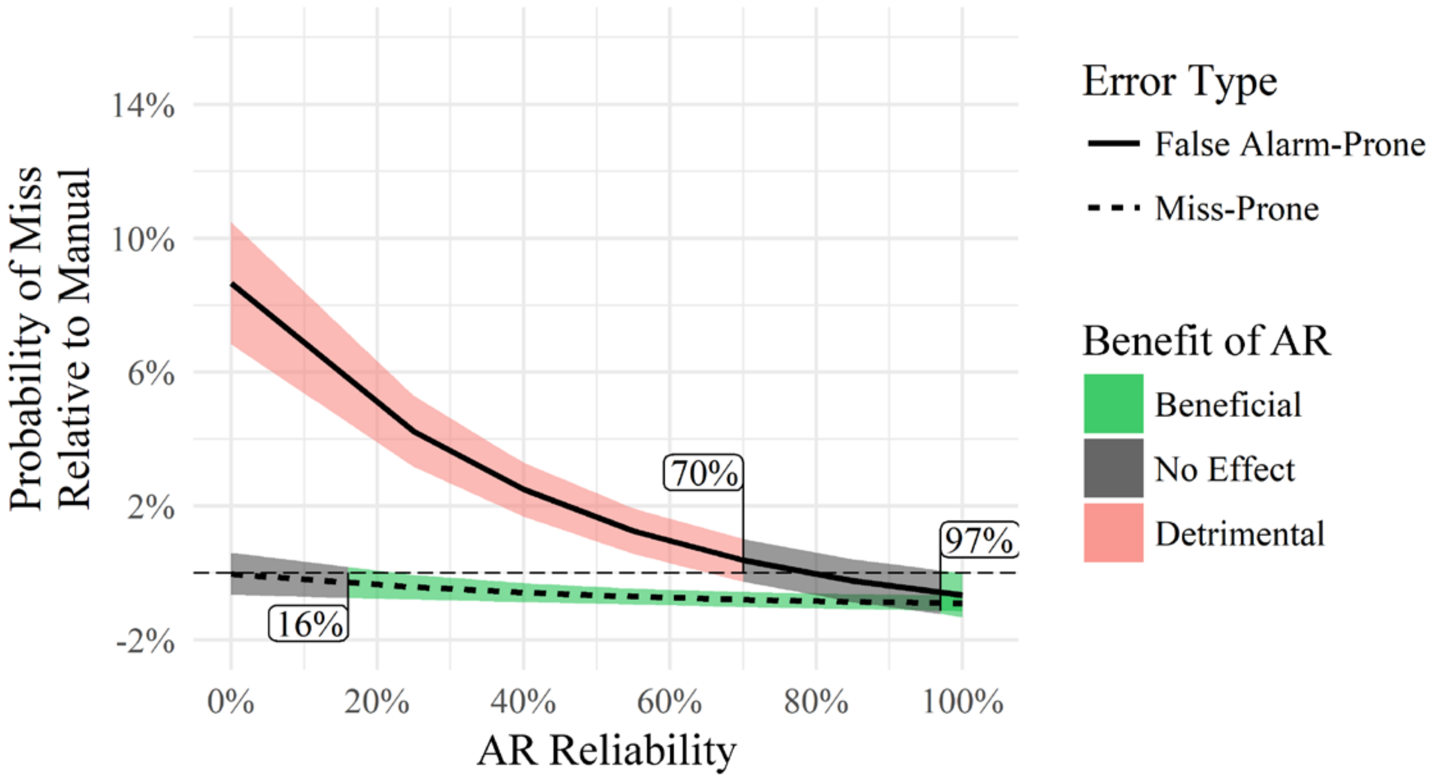
- **Tutorial and training**
  - Exposed to perfect and very unreliable AR
- **Baseline (no AR) performance**
  - Self-report survey asking about participants' se
    without AR
- **Instructed to use and evaluate 6 differen**
  - Self-report surveys administered after each of
- **All data were subtracted from no-AR bas**
  - Results represent the *change in performance*



PROBABILITY OF A MISS

- **Participants missed targets when paired with false alarm-prone AR**
  – Visual field cluttered with AR-marked targets: participants missed valid targets

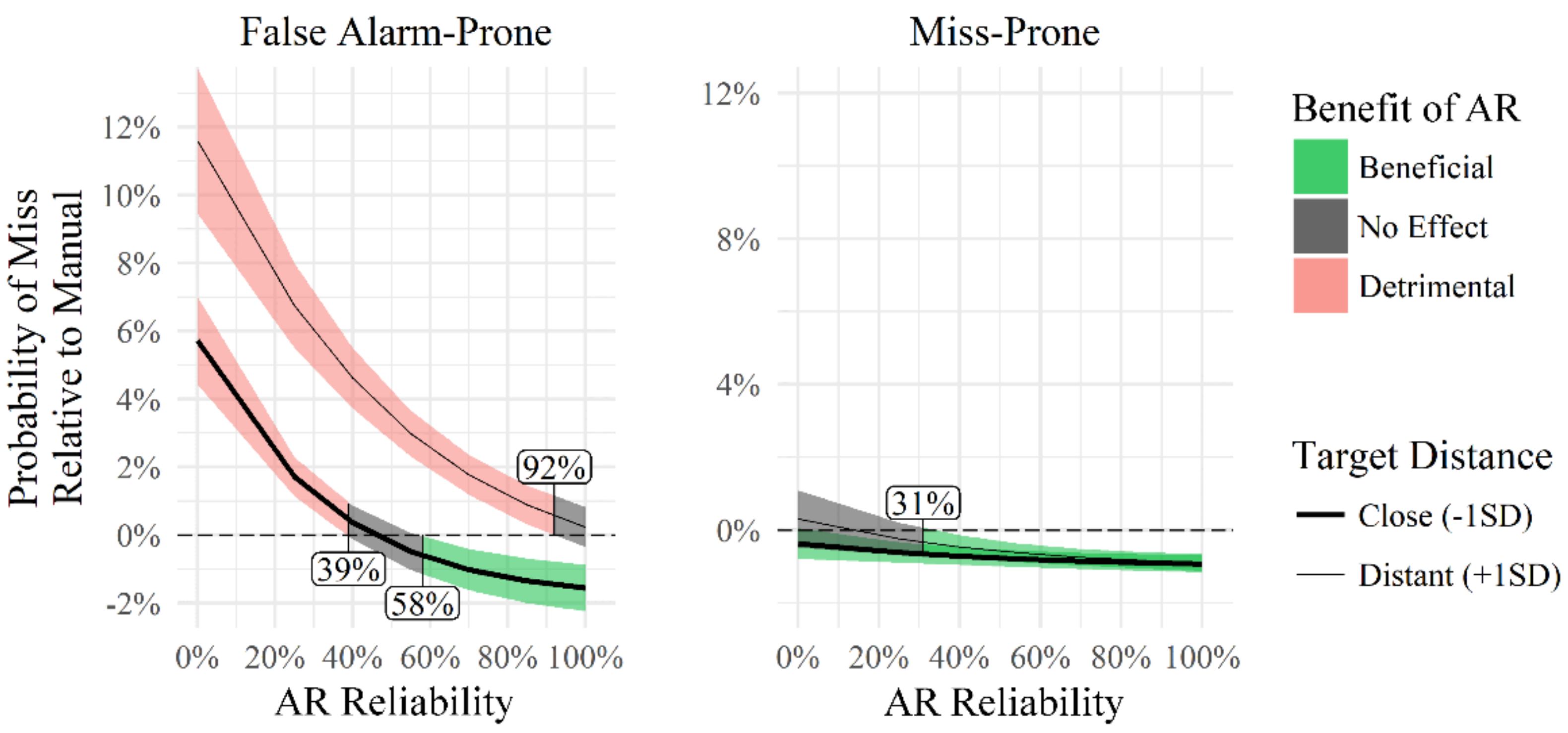
- **Miss-prone AR never hurt performance**
  – Visual field missing AR-marked targets: participants nonetheless found valid targets



PROBABILITY OF A MISS & RANGE

- **Distant targets were more difficult to discern: participants missed more of them**

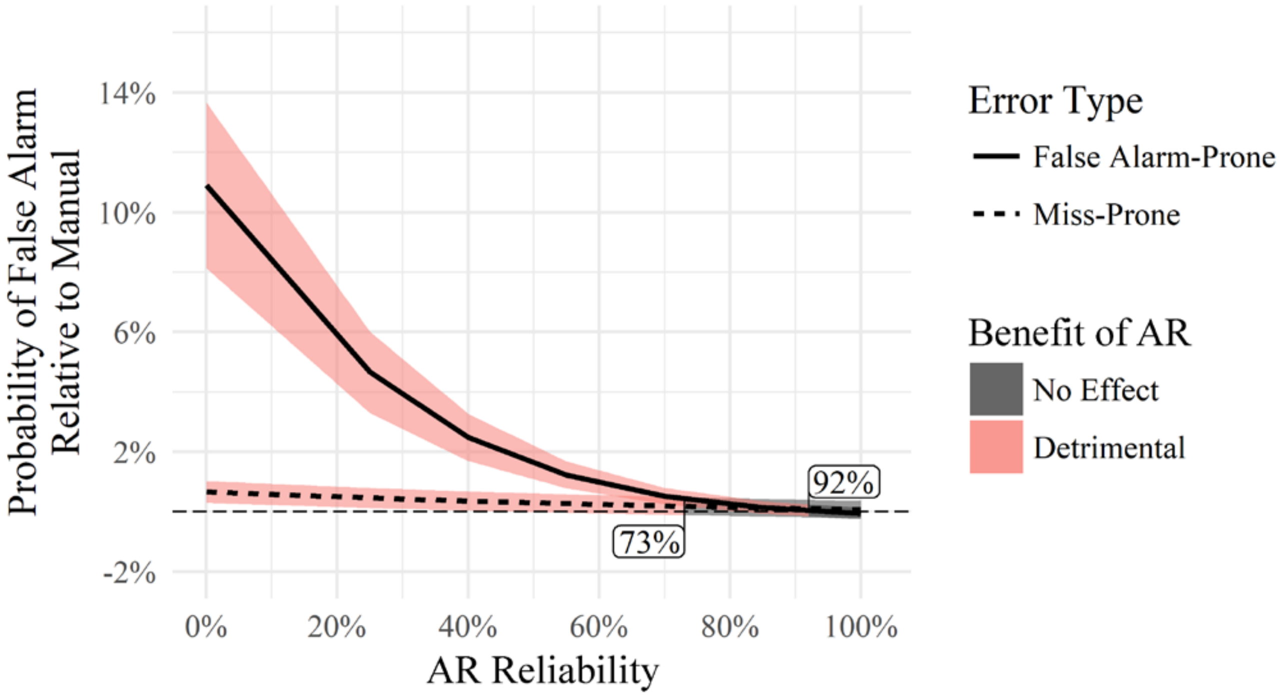
- **Distant targets magnified the undesirable effects of unreliable AR**
  - Distant targets even *more* likely to be missed with false alarm-prone AR (despite always being properly marked)



PROBABILITY OF A FALSE ALARM

- **Participants incorrectly selected non-targets when paired with false alarm-prone AR**
  – Visual field cluttered with AR-marked targets: participants selected invalid targets

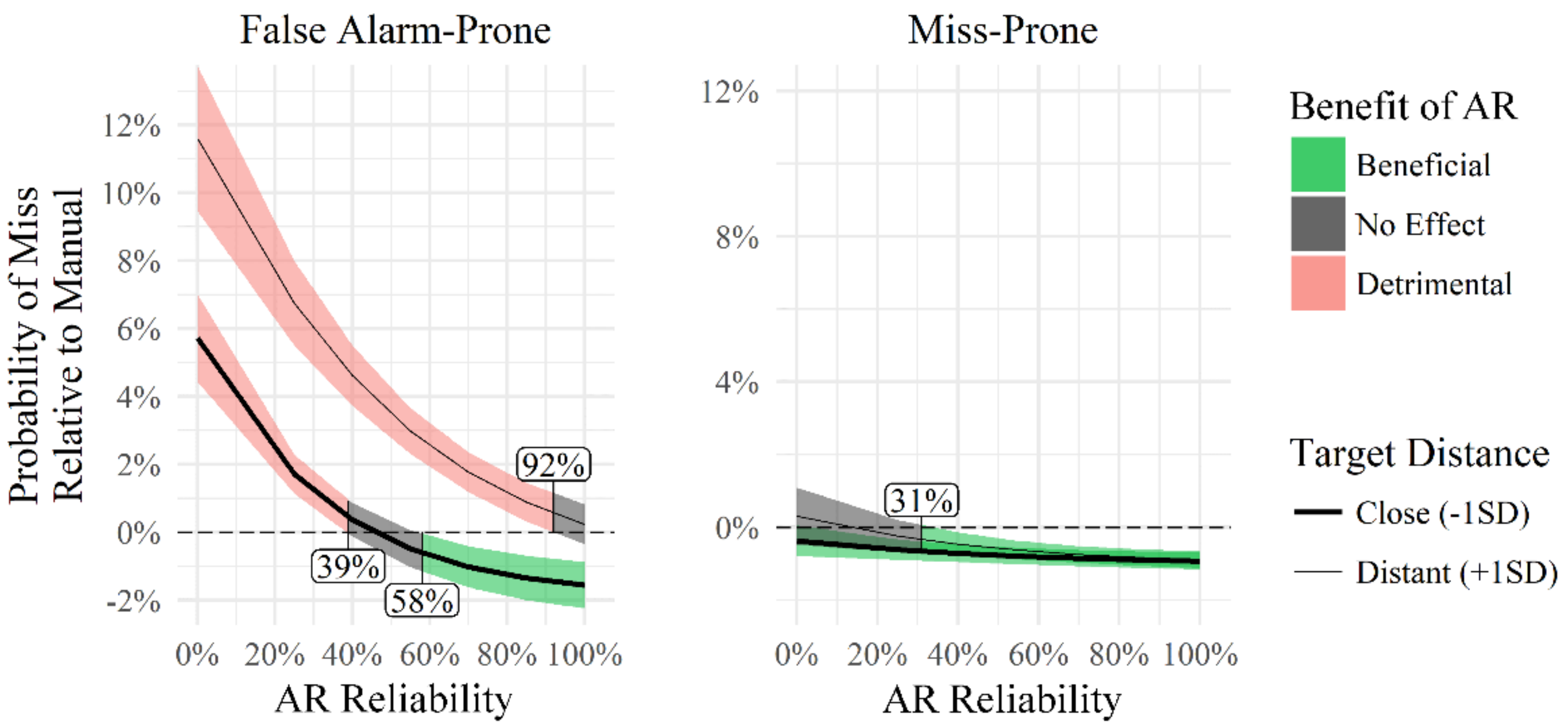
- **Miss-prone AR never hurt performance**
  – Visual field missing AR-marked targets: participants were not tempted to select invalid targets



PROBABILITY OF A FALSE ALARM & RANGE

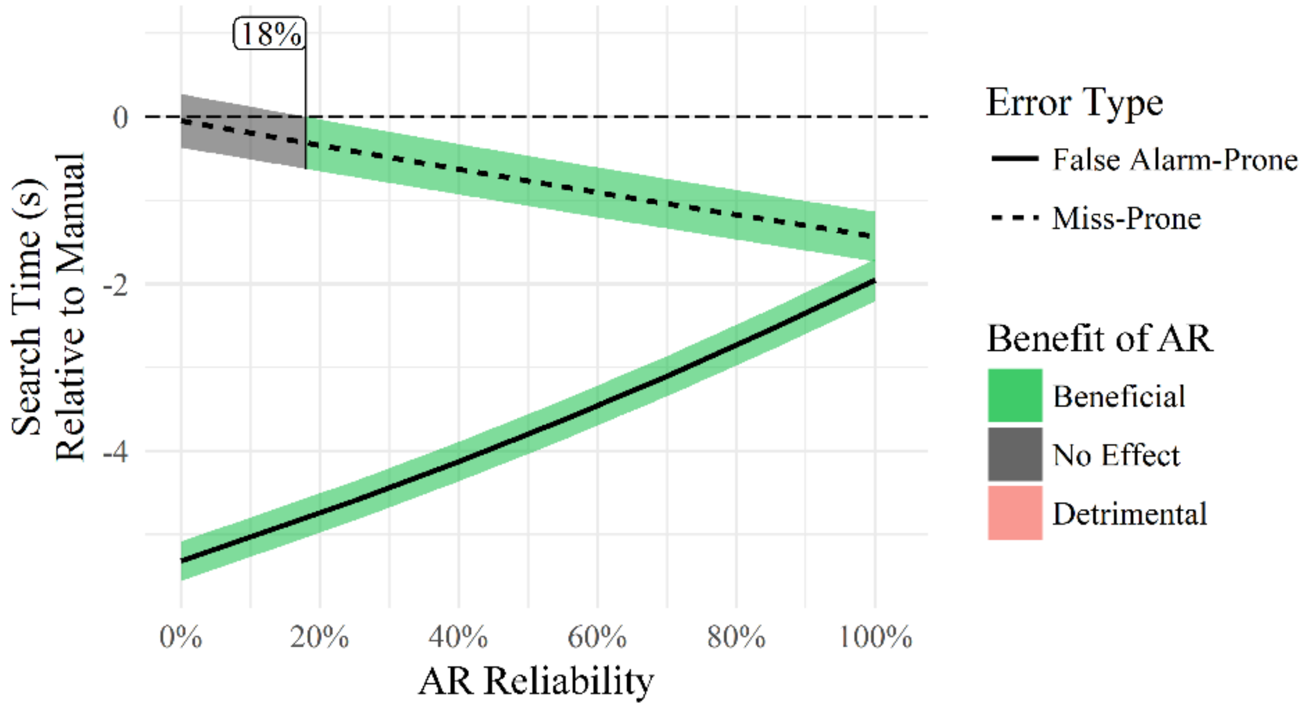
- **Distant targets were more difficult to discern, and elicited more false alarms than close ones**

- **Distant targets again magnified the effects of unreliable AR**
  – Erroneously-marked distant targets were *much* more likely to elicit false-alarms compared to close targets



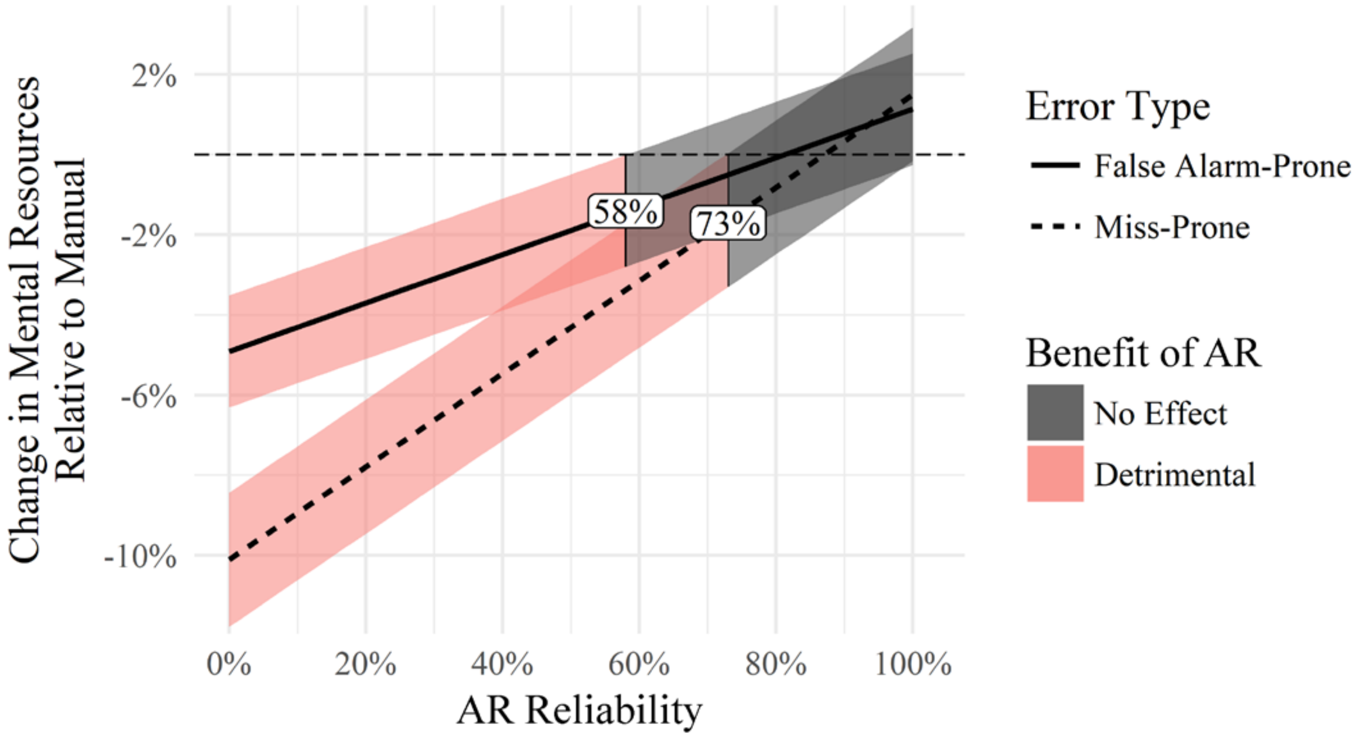
SEARCH TIME

- **Miss-prone AR: participants increased their search time as AR mistakes increased (adaptive)**
- **False-alarm prone AR: participants reduced their search time as AR mistakes increased (maladaptive)**
  – Less diligence/gave-up (but still made more responses)



SUBJECTIVE RESPONSES

- **Participants reported greater resource drain with miss-prone AR**
  - Coupled with longer search time: potentially working harder (i.e., not giving up)
- **Trust surveys and self-reported workload were similar between AR types**
  - Participants subjectively unaware of differences between conditions

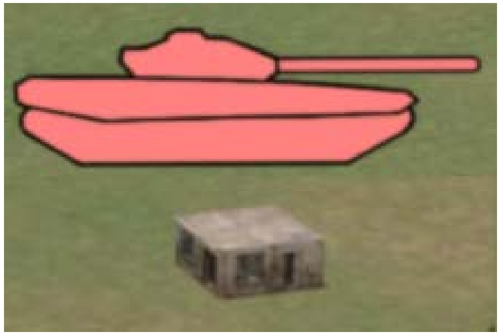


DISCUSSION: KEY FINDINGS

- **False alarm-prone AR was more damaging to accuracy**
  - Increased both misses and false alarms
- **Participants with miss-prone AR could compensate for poor AR by increasing effort**
  - False alarm-prone AR overwhelmed participants, who responded with less effort
- **Similar trust and workload self-report, despite objective performance differences**
  - Danger: in some cases, participants are unaware of when AR may hurt

DISCUSSION: AR ERROR TYPE

- We spend a great deal of effort avoiding “misses,” but participant misses increased with false alarm-prone AR
- Findings were consistent with prior research: false alarms can be more damaging at the same level of performance, are annoying and distracting
- Participants were unable or unwilling to pay the cognitive cost of working with false alarm-prone AR: more difficult task
- Even with highly motivated soldiers, perseverance may cost greater mental effort and result in inevitable mistakes

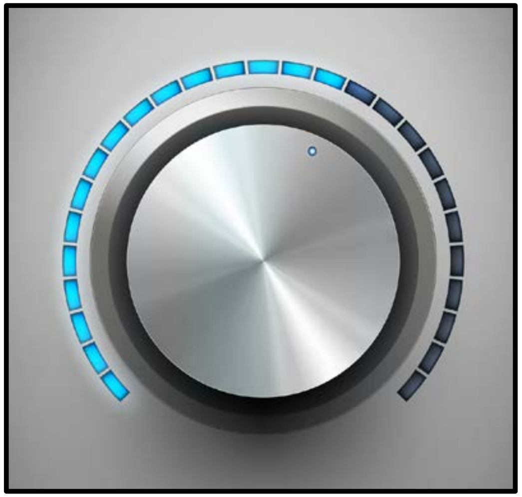


DISCUSSION: SENSITIVITY

- **Future AR systems may allow users to adjust detection thresholds (i.e., sensitivity)**
  - Knowing that misses and false alarms are not equivalent, how much freedom do we give users?

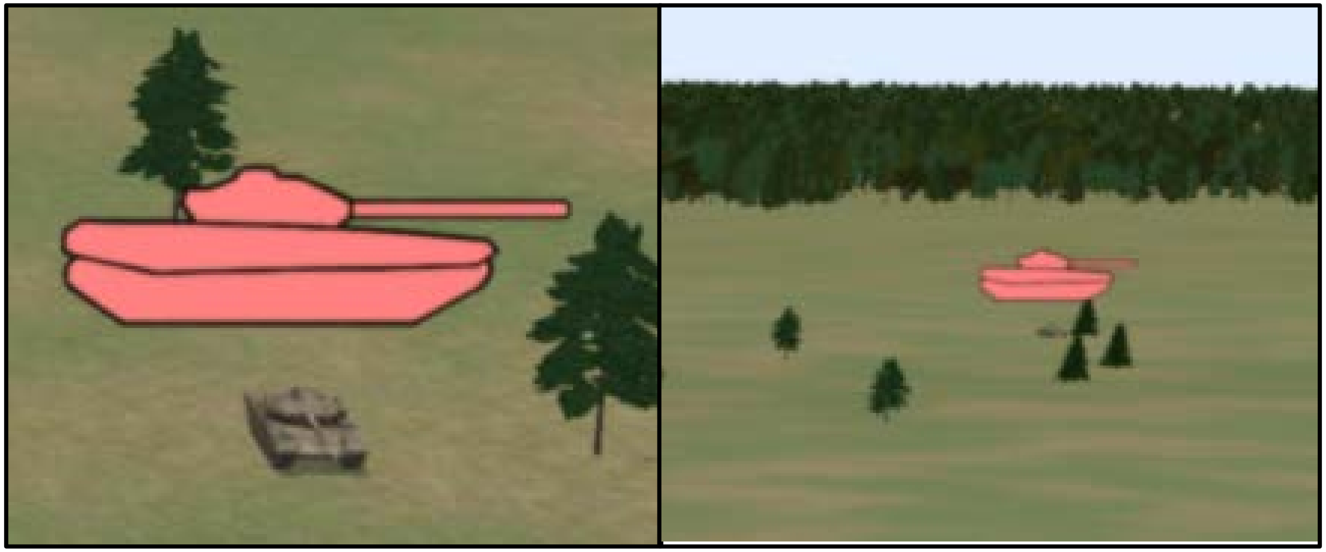
- **Users rated trust similarly, despite differences in objective performance – soldiers may not recognize risk of false alarms, especially if misses are "high cost"**

- **Potential solution: employ system constraints and/or user training to prevent alert oversaturation and disengagement from false alarms**



DISCUSSION: TARGET DISTANCE

- **Effects of unreliable AR on participant performance were magnified with distance**
  – At range, AR is providing more support; human cannot compensate
- **AR systems aiding target search will require greater reliability at longer ranges to improve human performance**
  – This finding may generalize to more difficult perceptual tasks in general
- **Do not automate/augment with insufficient accuracy – only automate/augment what you can do well**



CONCLUSION

- **AR accuracy required to improve human performance depends on contextual factors (AR error type, target range, etc.)**

- **In visual search, false alarms are more damaging to performance than misses are**

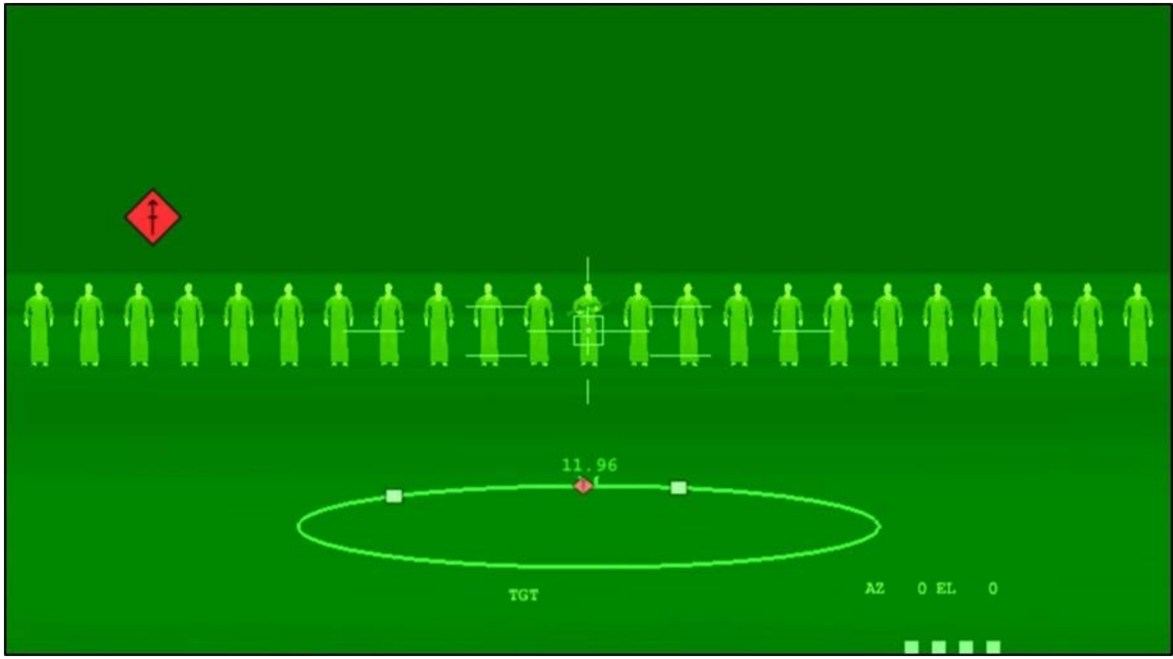
- **Disparity between objective performance and subjective responses suggests potential risk**
  - "We can't compensate for poor AR if we don't know that it's hurting us"

- **AR will rarely be perfect, but <u>it should improve human performance over a "manual performance" baseline</u>**

NVESD Perception Laboratory Augmented Reality Program Overview

QUALITY OF AR INFORMATION

- **NVESD interests: Sensor feeds, see-through displays**
  - Many potential benefits to AR technology for Soldiers (situational awareness, decision-making, communication, etc.)
- **Assumption: providing Soldiers with AR information will improve their performance**
- **Many factors affect quality of AR information**
  - Perceivable?
  - Intuitive?
  - Timely?
  - Relevant?
  - Accurate?
- **Initial research areas:**
  - Visual Search,
  - Target Acquisition,
  - Vehicle Identification,
  - Navigation



AR RED TEAM RESEARCH OBJECTIVES

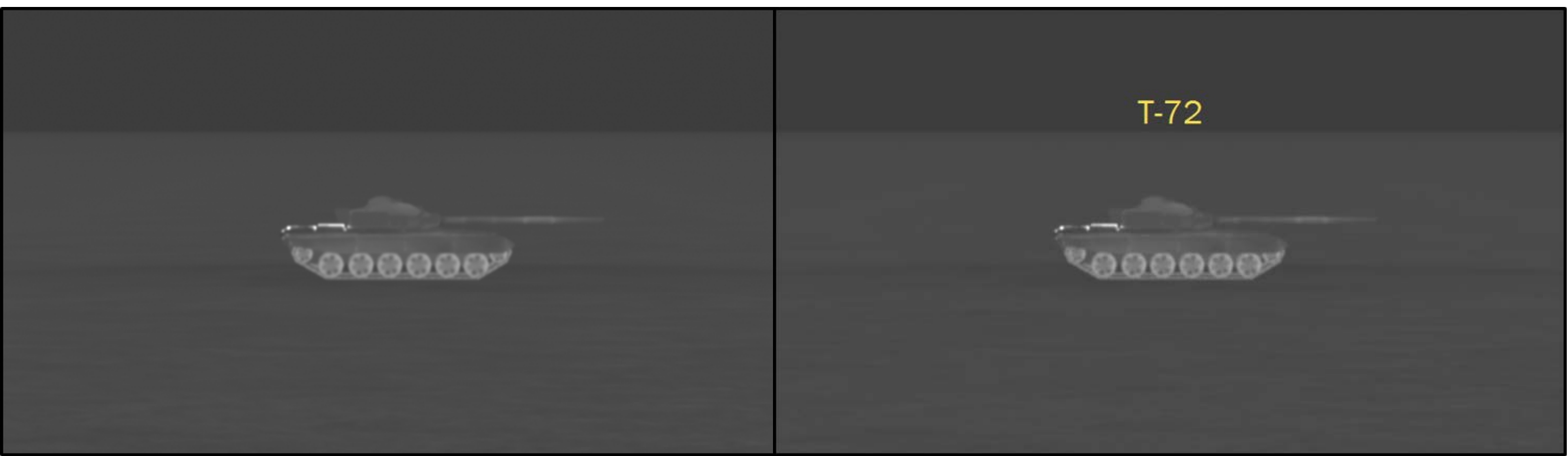
Our simulations currently focus on AR accuracy and human performance:

- How accurate does AR have to be in order to improve performance?
- What are the worst types of errors an AR system can make?

These are task-specific and potentially device-specific questions

Goals:

1) Contribute to general AR usage guidelines
2) Adapt our existing simulation capabilities to be able to define sensor- and task-specific AR requirements



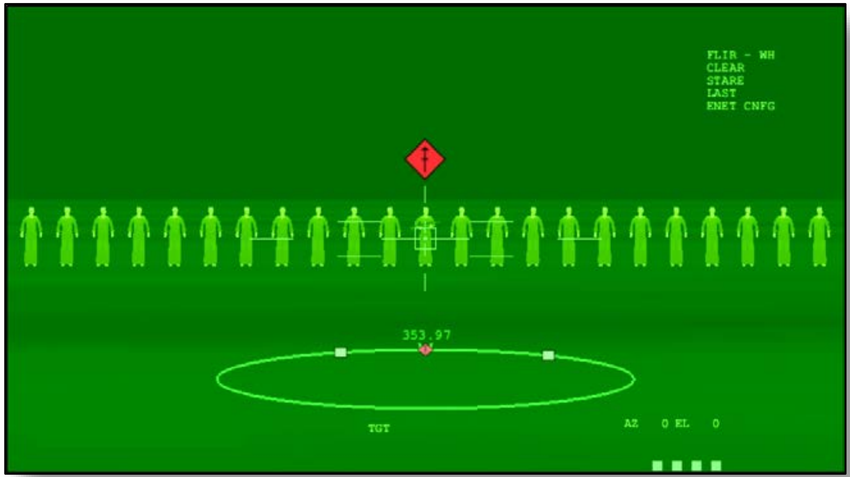
Target Acquisition and Spatial Error

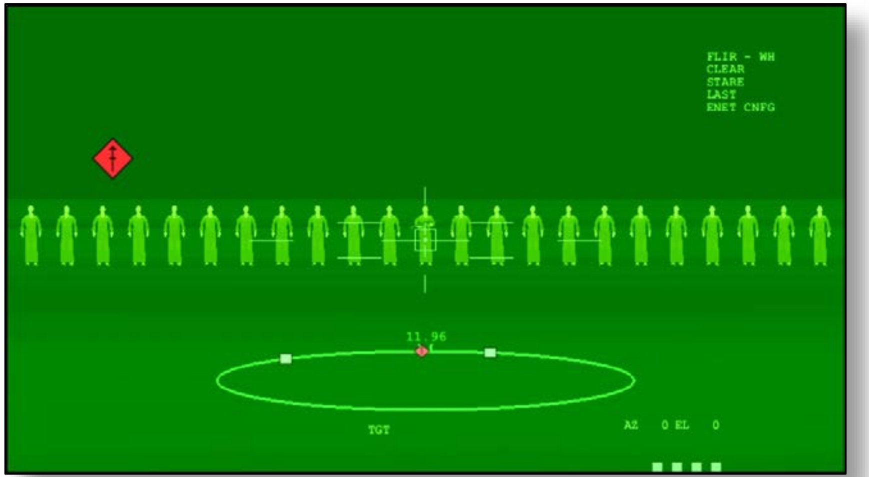
TARGET ACQUISITION

<u>Research Objectives:</u>

- Evaluate AR aid to target acquisition performance and how errors impede performance
- Evaluate the level of AR accuracy necessary to improve target acquisition performance at various ranges

Example imagery depicting a scenario with an AR designation perfectly aligned (left) and misaligned (right)





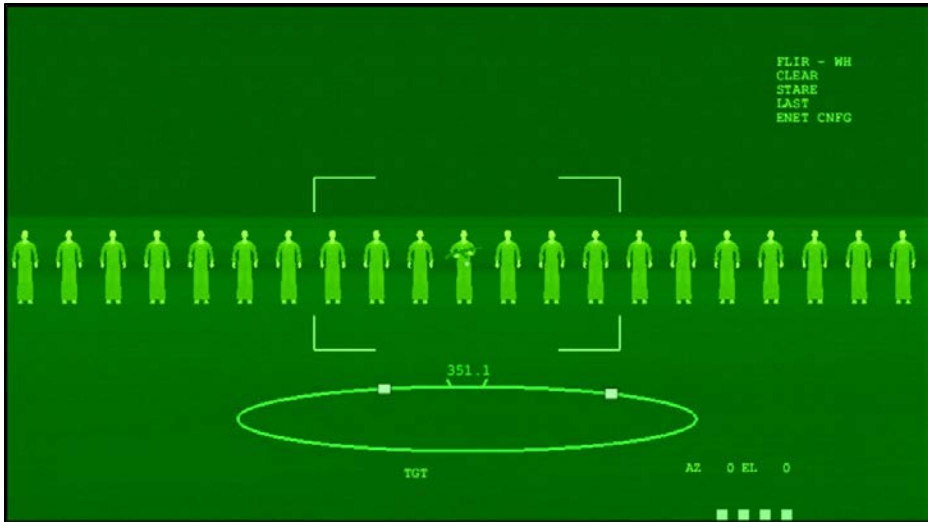
METHODOLOGY

Scene Generation in Night Vision Image Generator Software (NVIG):

- Virtual humans arranged in a 60˚ arc around the sensor, placed closely together (1m apart)

- A single target held an AK-47

- Participants: 18 U.S. Soldiers



EXPERIMENTAL DESIGN

- 6 AR Conditions: No AR, Perfect AR, plus 1°, 2°, 3°, and 4° of angular error

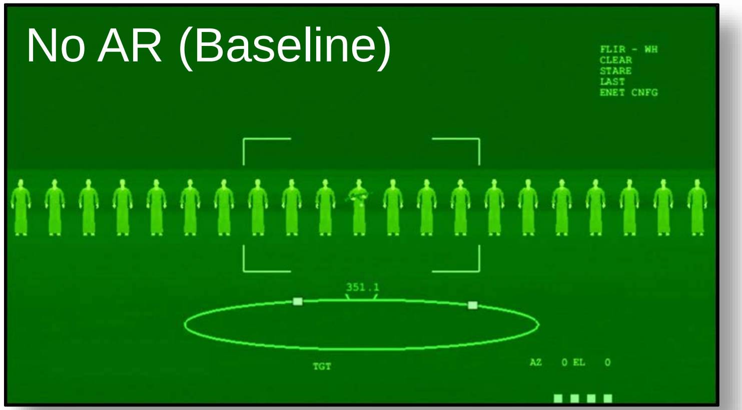
- 3 Ranges: “Close,” “Intermediate,” and “Distant”

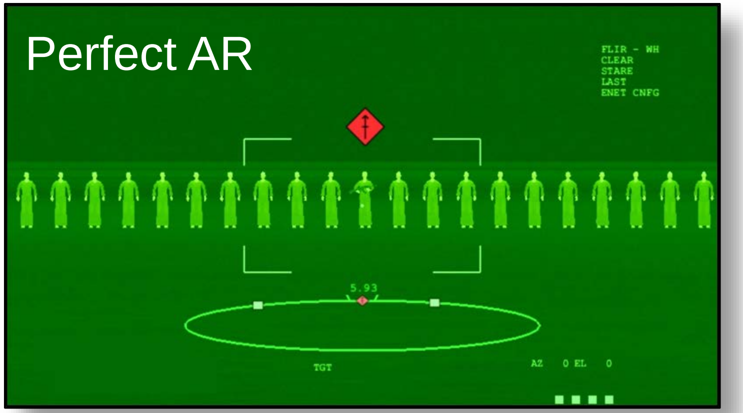
- Targets placed randomly within 3° sections, centered at 6°, 9°, 12°, and 15°
    - Target locations were counterbalanced across all AR Condition and Range combinations

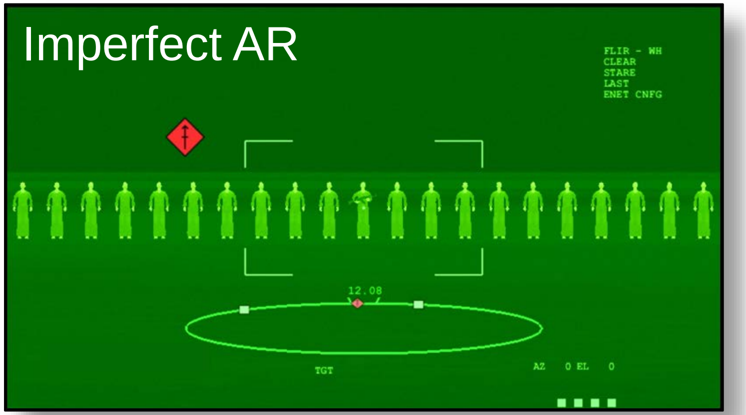
- 144 trials, divided in 8 blocks (rest)

- Counterbalanced the 8 blocks of trials by AR Condition and Range

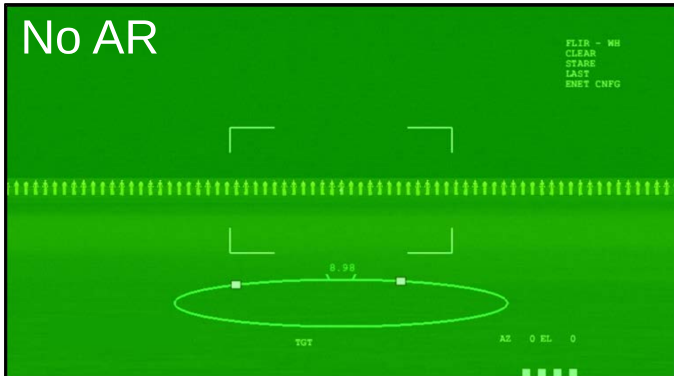
“CLOSE” RANGE

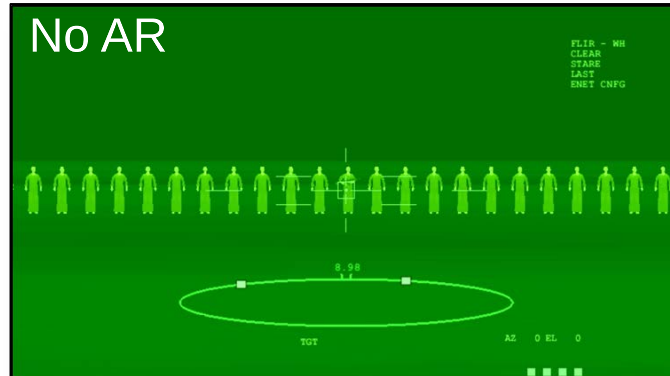

No AR (Baseline)
FLIR - WH
CLEAR
STARE
LAST
ENET CNFG
351.1
TGT
AZ 0 EL 0


Perfect AR
FLIR - WH
CLEAR
STARE
LAST
ENET CNFG
5.93
TGT
AZ 0 EL 0


Imperfect AR
FLIR - WH
CLEAR
STARE
LAST
ENET CNFG
12.08
TGT
AZ 0 EL 0

"INTERMEDIATE" RANGE

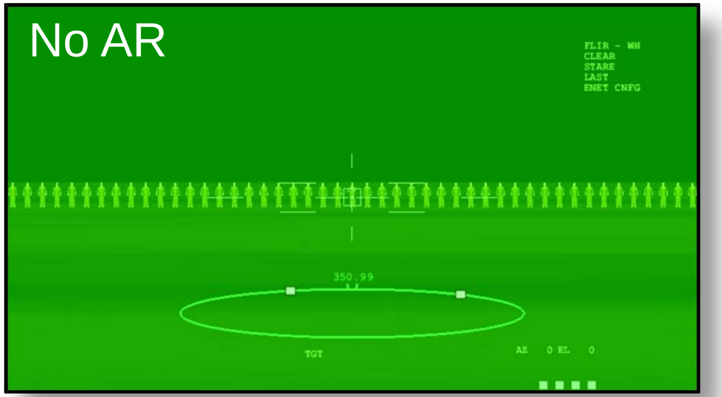


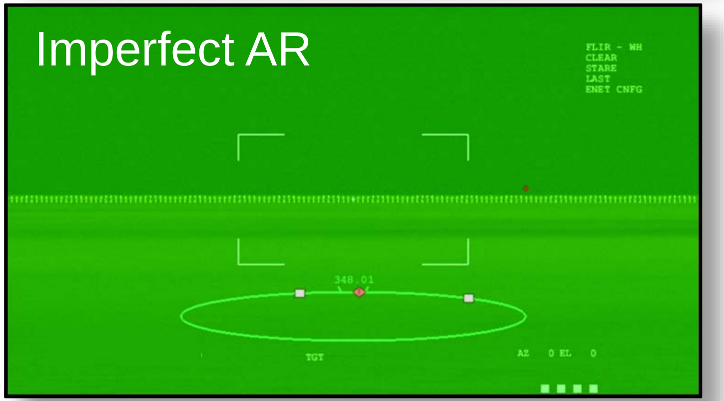


No Optical Zoom

Optical Zoom

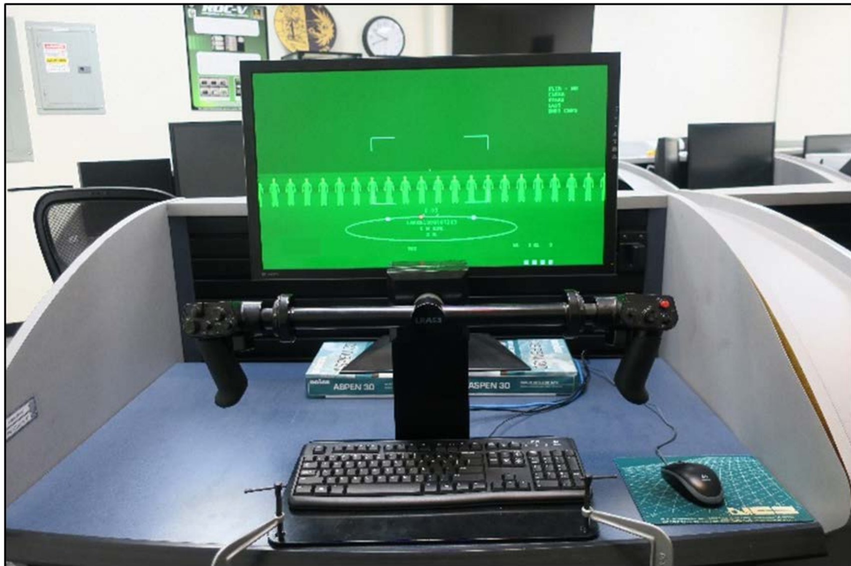
“DISTANT” RANGE





PROCEDURE

- Participants stayed for 5 days (vehicle identification training, other experiments); two cohorts
- Highly realistic sensor grips – simplified controls for optical zoom, “speed boost,” and target designation
- Group presentation on experiment and controls
- 27 training trials (3 trials each of No AR, Perfect AR, and 4° of angular error at each range)
- Experiment: breaks as desired between blocks of trials,10 minutes at halfway point
- Length: ~120 minutes



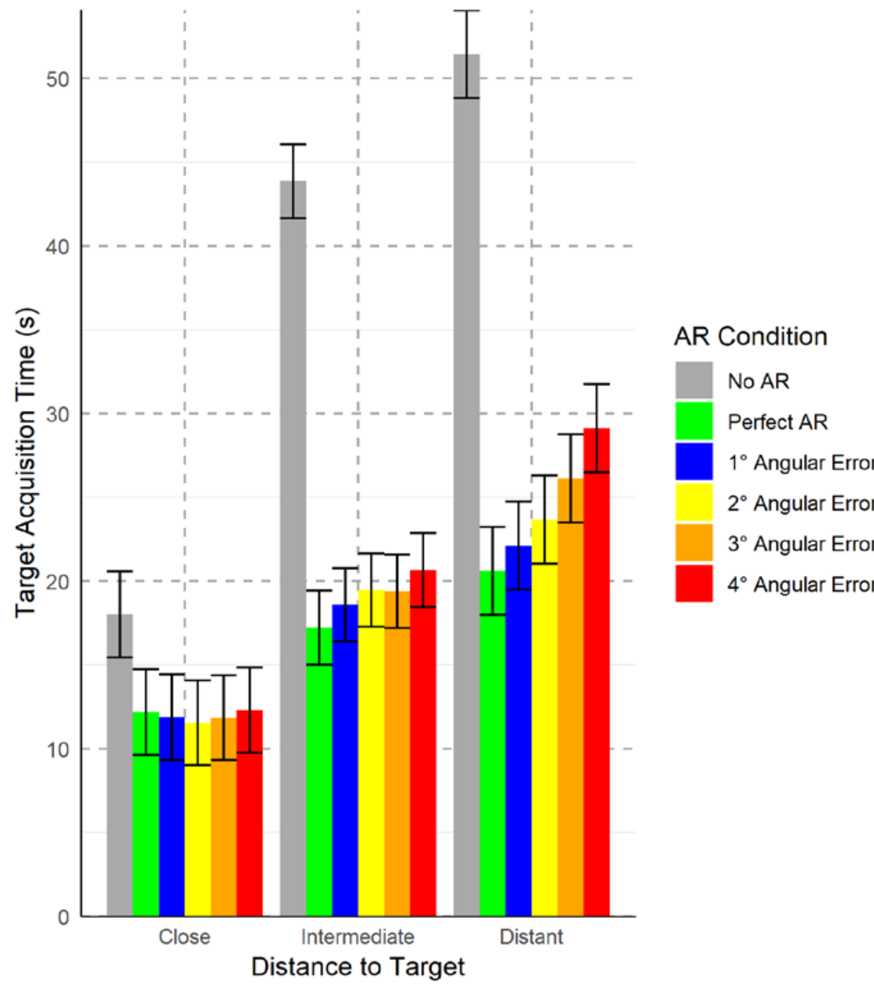
RESULTS: TARGET ACQUISITION TIME

Significant Main Effects:

- Range
- AR Condition

Compared to No AR:

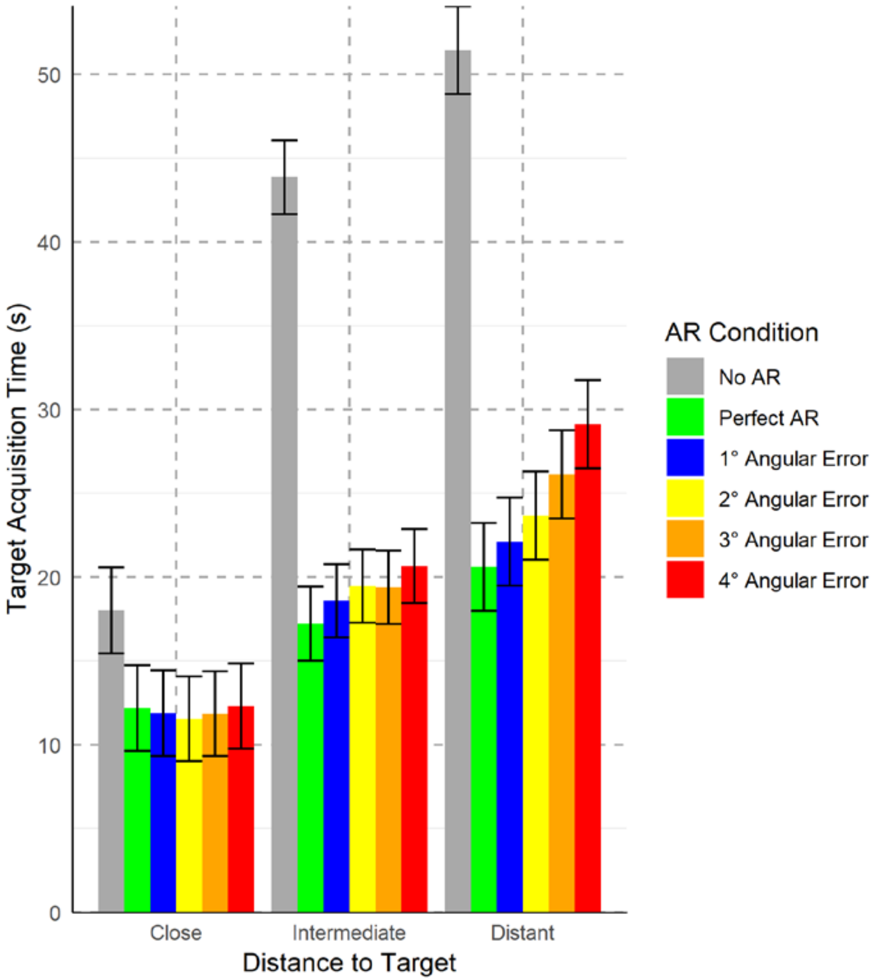
- All AR conditions improved performance
- All AR conditions protected against increased reaction time with increased range
  - Protection decreased w/angular error



RESULTS: TARGET ACQUISITION TIME

Compared to Perfect AR:

- 1° and 2° did not significantly impair performance, but 3° and 4° did
- 3° and 4° also showed greater increases as range increased
- No significant differences at “Close” range for any imperfect AR conditions
- 3° significantly worse at “Distant” range
- 4° significantly worse at “Intermediate” and “Distant” ranges



DISCUSSION

Results Summary:

- Incremental degradations in AR accuracy produced progressive degradations in performance
- Even imperfect AR was always beneficial in this simulation (won’t be true for every task)
- Greater AR accuracy is needed at greater ranges: all AR yielded approximately the same benefit at the “Close” range, but greater error at the “Distant” range yielded deficits compared to perfect AR

Vehicle Identification Accuracy

VEHICLE IDENTIFICATION USING LWIR IMAGERY

Research Objectives:

- Evaluate AR aid to vehicle identification performance and how errors impede performance
- Evaluate the level of AR accuracy necessary to improve vehicle identification performance at various ranges

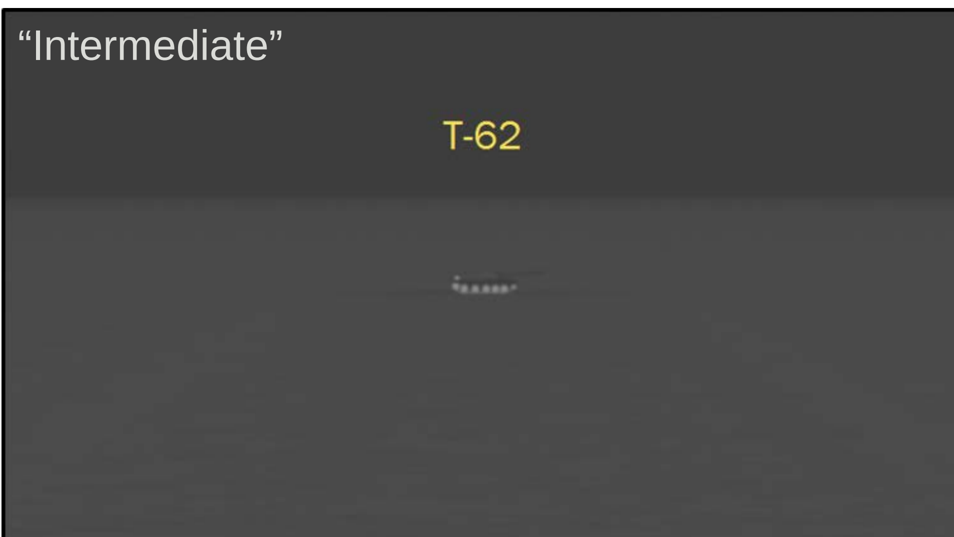
Correct Label	Incorrect Label

EXPERIMENTAL DESIGN

Independent variables

- AR Conditions: 100%, 75%, and 50% reliable AR, No AR pretest and posttest
- 3 Ranges
- Time to make a decision: unlimited time vs. 5 second time constraint

Dependent Variables: Accuracy & Response time



PROCEDURE

- **20 U.S. Army Soldiers – trained on infrared vehicle ID prior to experiment**
- **Scene Generation in NVIG**
- **Sequence: Training, No AR Pretest, AR Trials, No AR Posttest**
- **No AR and AR components had both a time-constrained and a time-unlimited portion**
  - All participants took both
  - Randomly assigned to always begin with time limit or no time limit
- **Baseline:** 72 Images (3 blocks of 24, 1 block per range)
- **AR Trials:** 216 images (9 blocks of 24, 1 block per range X AR reliability)
- **Participants were told to evaluate different ostensible AR systems**
  - After each block of images, asked to reset their trust
- **Participants asked to take breaks between sections of the test**

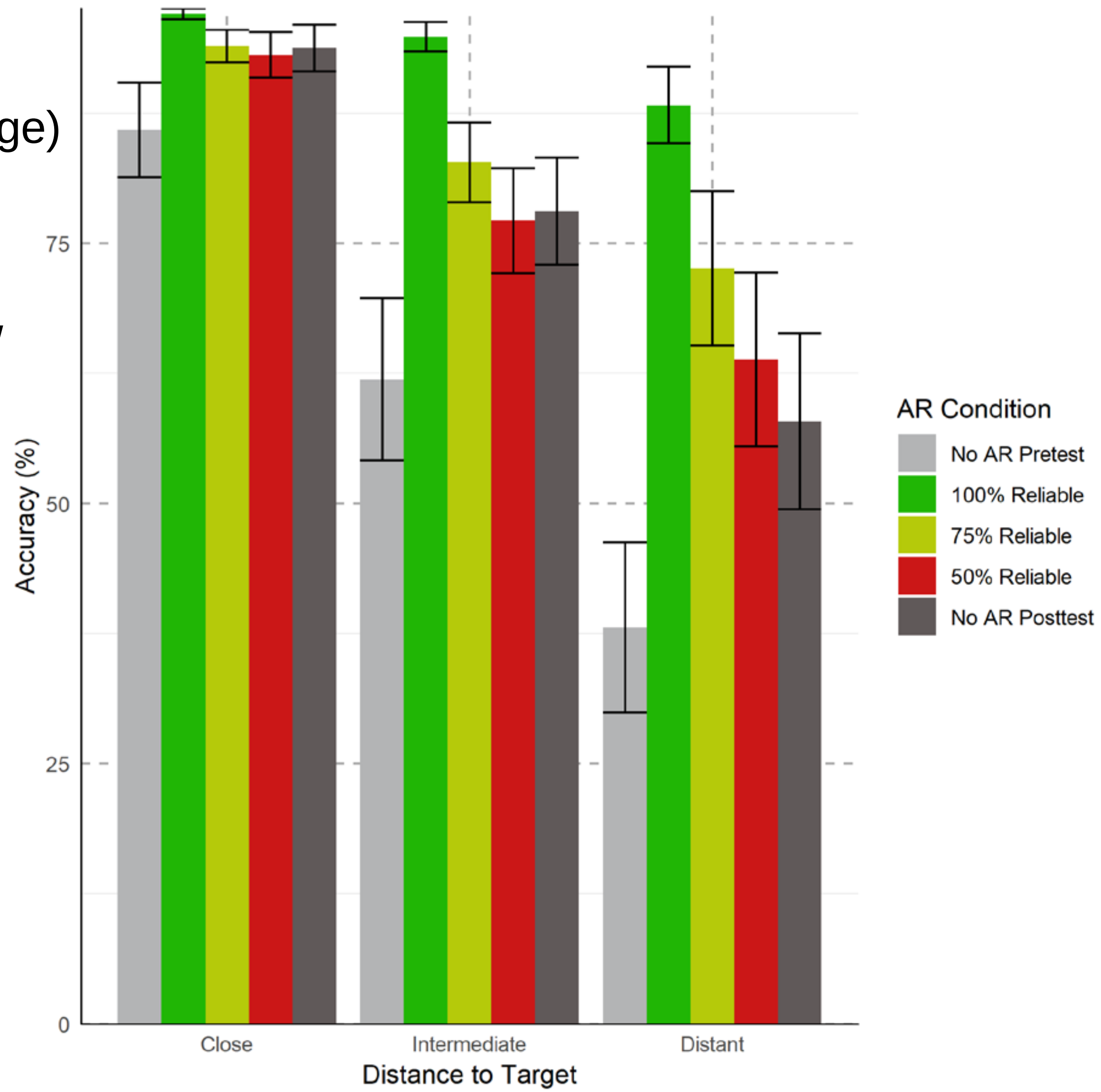
ACCURACY

Significant main effects

- AR Condition
- Distance to target (i.e., range)
- Time constraint

Substantial differences b/w pretest and posttest:

- Indicates substantial learning during experiment
- Posttest selected as reference



ACCURACY WITH AR COMPARED TO UNAIDED PERFORMANCE

"Close" Range

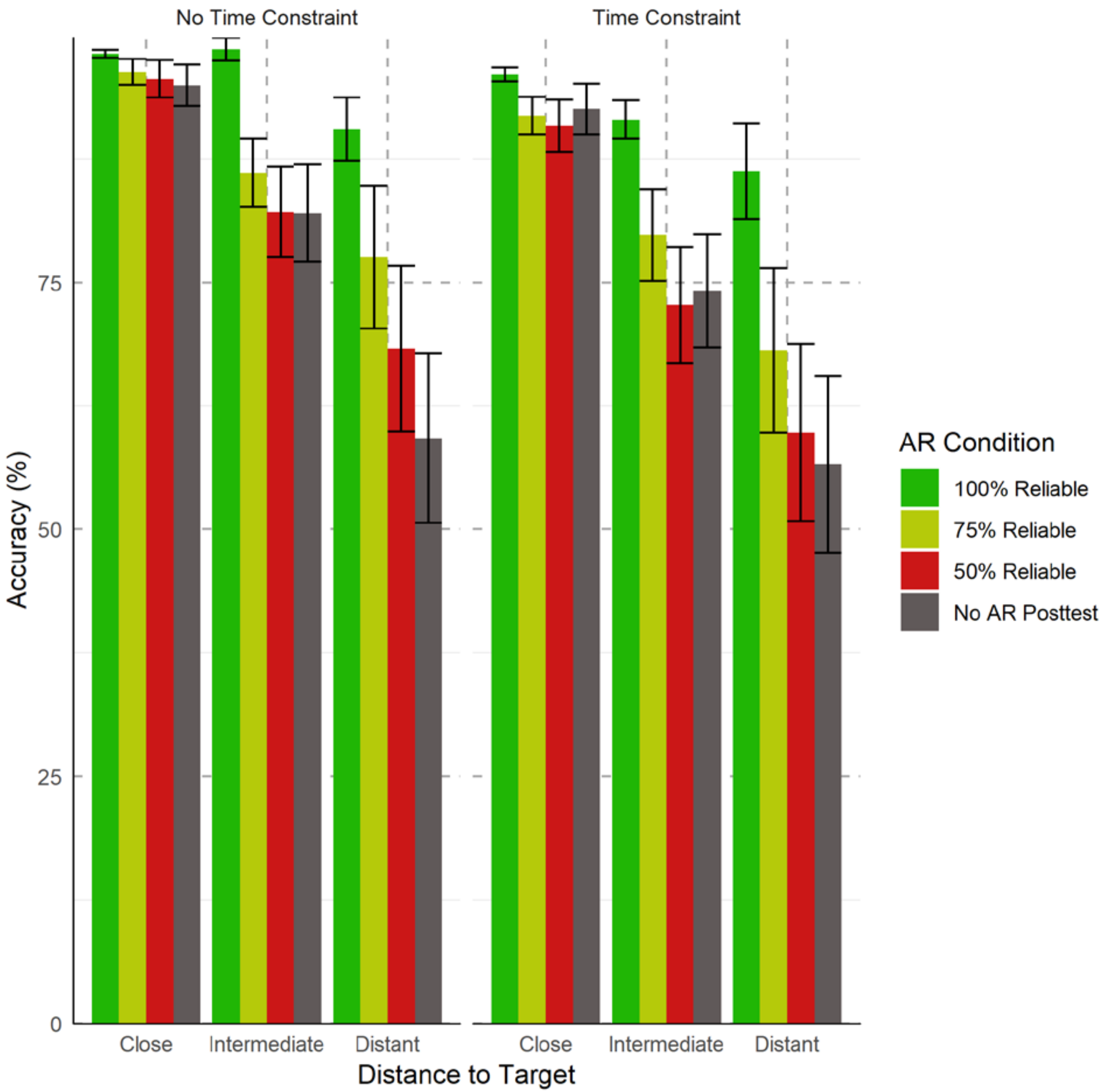
- Perfect AR approached significant improvement
- Imperfect AR: non-significant

"Intermediate" Range

- Perfect AR: significant improvement
- Imperfect AR: NS
- Time-constraints caused a greater reduction in perfect AR accuracy

"Distant" Range

- All AR information is a significant improvement



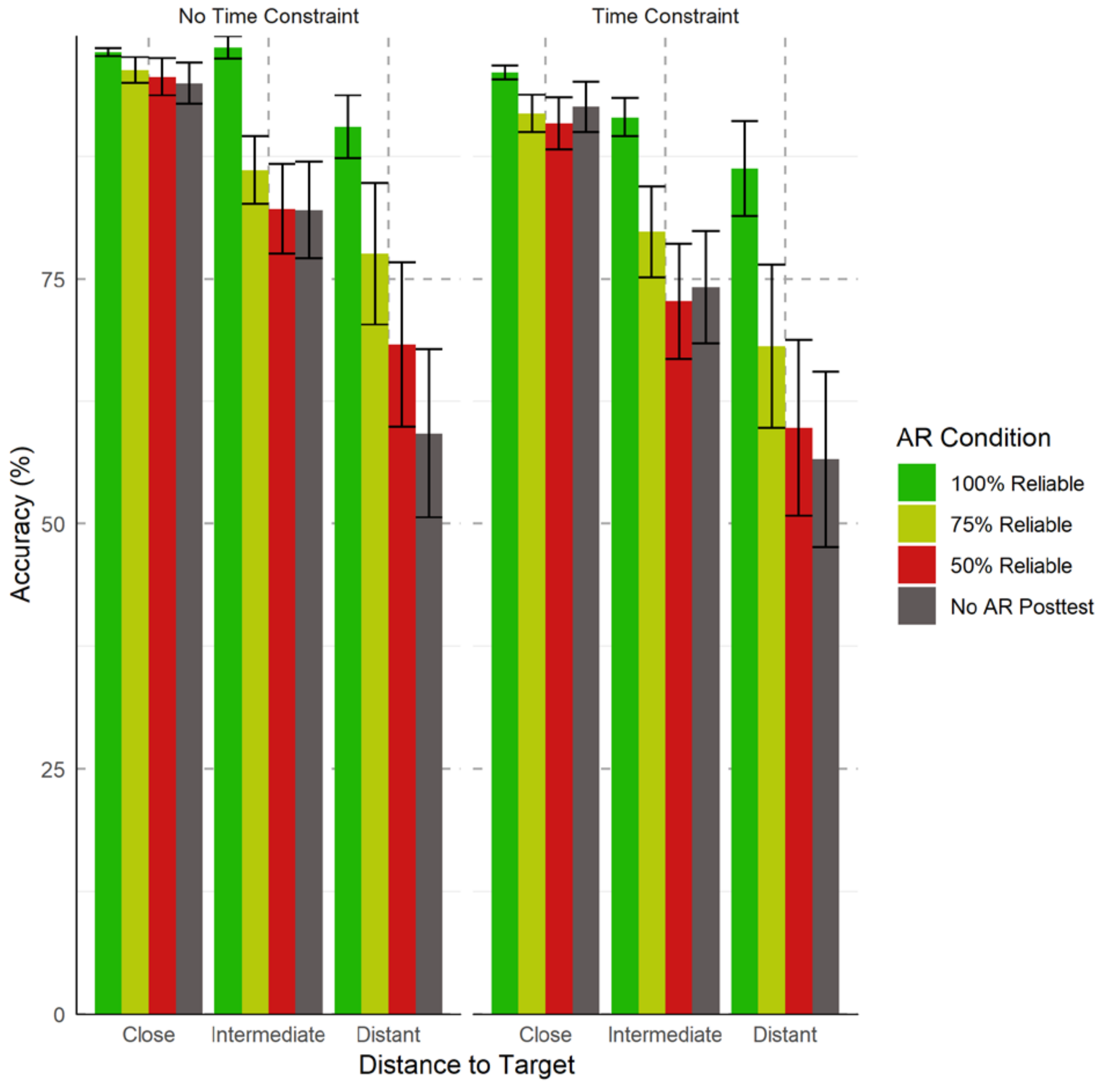
ACCURACY WITH IMPERFECT AR COMPARED TO PERFECT AR

"Close" Range

- 50% reliable AR caused significant impairments
- 75% reliable AR caused impairments that approached significance

Intermediate and Distant Ranges

- Both 50% and 75% reliable AR caused significant impairments

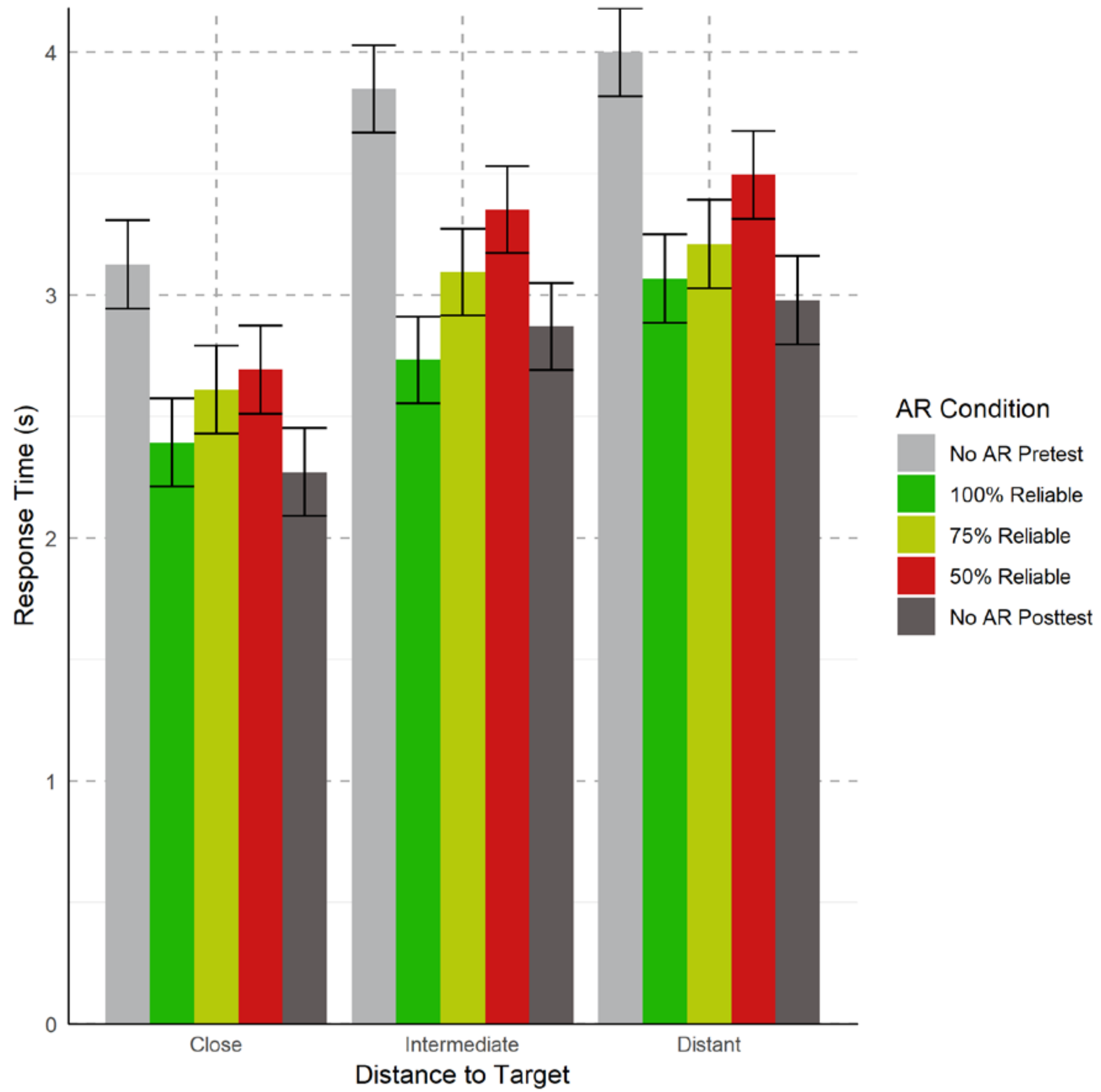


RESPONSE TIME

Significant main effects

- AR Condition
- Distance to target (i.e., range)
- Time constraint

- Substantial shift b/w pretest and posttest



RESPONSE TIME WITH AR COMPARED TO UNAIDED PERFORMANCE

“Close” Range

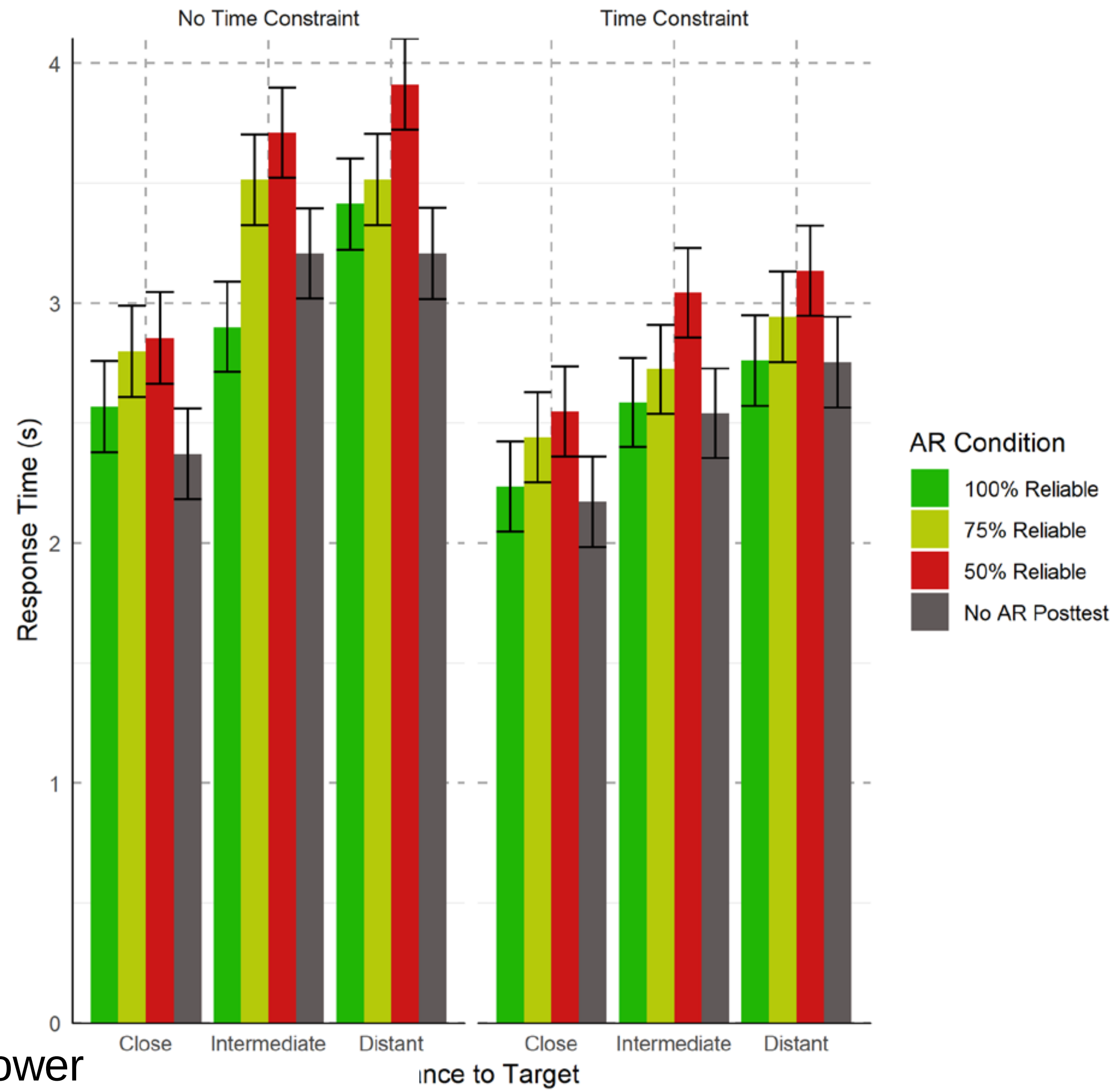
- 50% and 75% reliable AR – significantly slower
- Perfect AR – slower, approached significance

“Intermediate” Range

- 50% reliable: significantly slower
- 75% reliable: slower, approached significance
- Perfect AR: significantly faster
- Time-constraints: perfect AR decrease was not as severe

“Distant” Range

- 50% reliable – significantly slower
- 75% reliable and perfect AR – slower, but not significantly slower



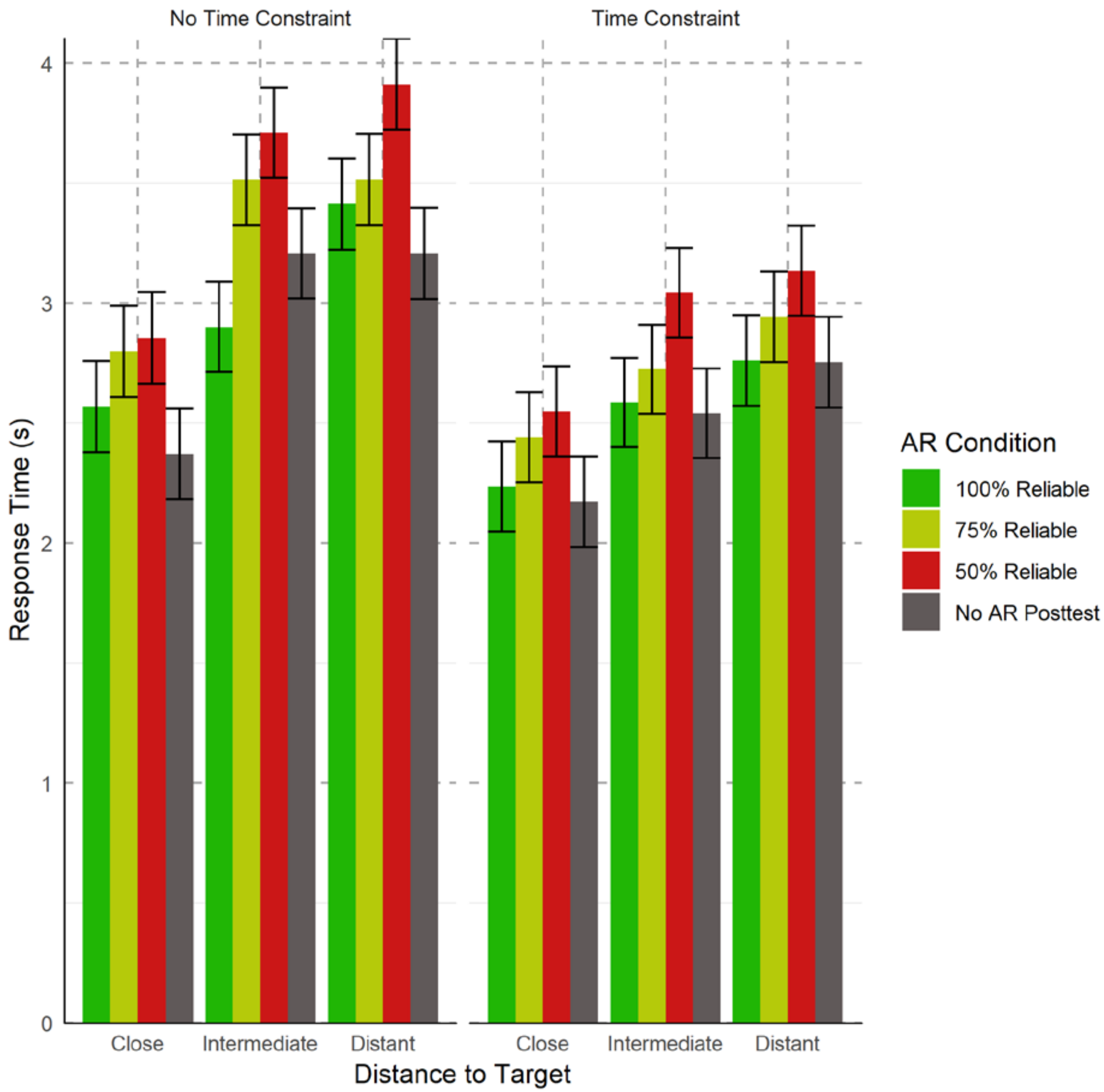
RESPONSE TIME WITH IMPERFECT AR COMPARED TO PERFECT AR

"Close" and "Intermediate Range

- Both 50% and 75% reliable AR: significantly slower

"Distant" Range

- 50% significantly slower
- 75% not significantly different



DISCUSSION: RESULTS SUMMARY

- **Progressive AR error yielded progressive impairments**
- **Participants were able to use perfect AR effectively**
- **Greatest benefit w/perfect AR, at "Distant" range, with unlimited time**
- **Imperfect AR was only clearly beneficial at "Distant" range when participants were clearly struggling**
  - 50% reliable AR always slower compared to No AR and perfect AR
  - 75% reliable AR showed many similar impairments, just less severe
- **Trends by range:**
  - Close: little AR benefit, yet slowed participants down
  - Intermediate: Perfect AR clearly beneficial, but time-constraints hurt improvement more than other conditions
  - Distant: All AR beneficial, greater reliance on AR
- **Time Constraints**
  - Significantly impaired performance
  - Generally affected AR conditions similarly (except "Intermediate w/perfect AR)
- **Most benefits to Accuracy** – usually a relatively small cost of speed

DISCUSSION: LIMITATIONS & FUTURE WORK

- **Experimental design: No AR trials were separate pretest and posttest**
  - Designed intentionally to capture learning/practice effects and to reduce length of core experimental trials
  - Became disadvantageous with severe learning effects
  - May overestimate baseline performance compared to other conditions
  - Future iterations: additional initial practice with NVIG simulated imagery and integrate baseline with other trials

- **Participants in our study expect AR mistakes**
  - May have caused additional skepticism with perfect AR (slower responses)
  - Results may not generalize to unexpected AR errors

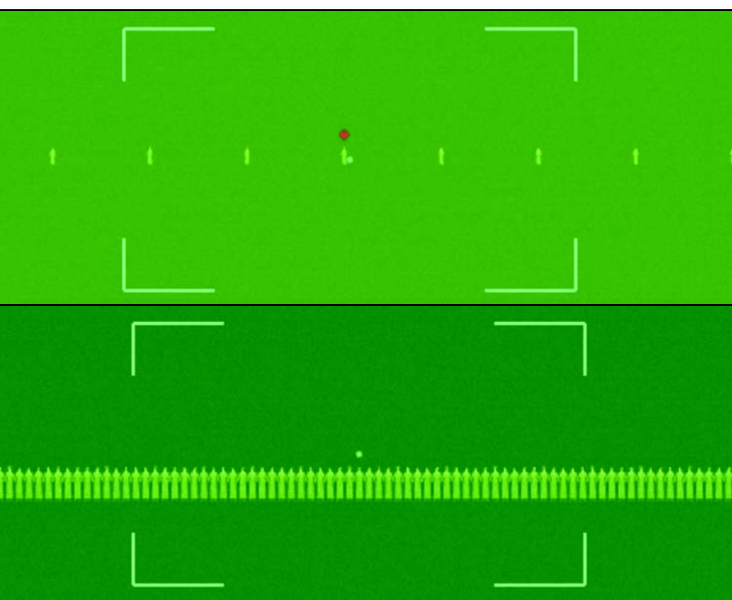
- **Broad AR reliability intervals (25%)**

Future Efforts

ONGOING AND FUTURE PROJECTS

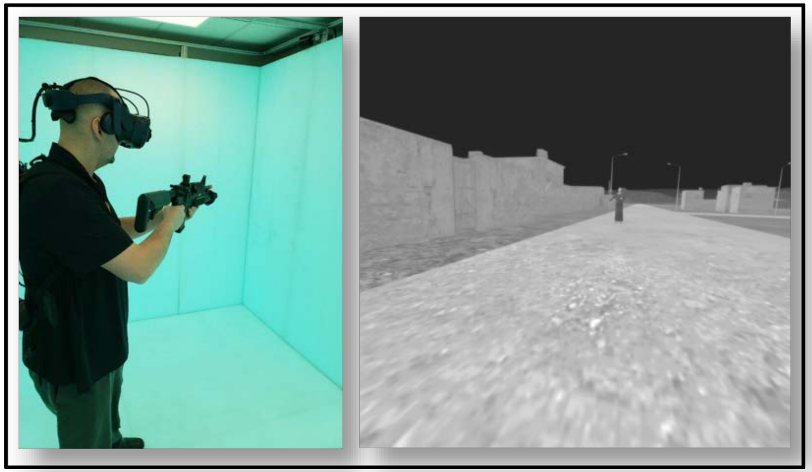
DEVCOM CSISR CENTER

- **Target Acquisition:** target density, clutter, field of view, & field of regard
- **Vehicle Identification:** Algorithms biased towards threats and other imagery degradations
- **Visual Search:** Misses and false alarms in a high clutter environment
- **Land navigation:** imperfect waypoints

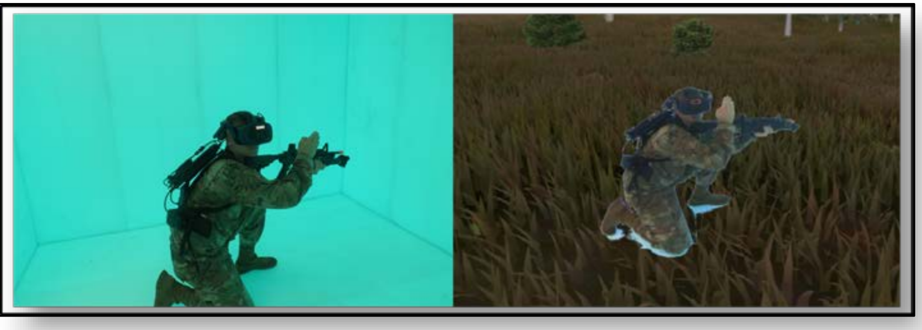


ONGOING AND FUTURE PROJECTS

- Immersive AR Display simulations – conducted in NVESD's mixed-reality Virtual Prototyping Holodeck (VPH)
- Target acquisition in an immersive environment
- Eye-tracking studies examining effectiveness and efficiency of AR Symbology



A person in the VPH (left) and the scene he sees in his VR display (infrared scene rendered by NVIG)



A Soldier gives a hand signal (left) and how he appears to a fellow Soldier in the VPH (Right)

QUESTIONS?