Study Protocol: Human-AI Collaboration Tester (HAICT) Exp. 7

ClinicalTrials.gov ID NCT05272189

Date December 3, 2025

- Study Protocol: The written description of the clinical study, including objective(s), design, and methods. It may also include relevant scientific background and statistical considerations (if the protocol document includes the statistical analysis plan, use "Study Protocol with SAP and/or ICF" option). Note: All amendments approved by a human subjects protection review board (if applicable), before the time of submission and that apply to all clinical trial Facility Locations must be included.
- Statistical Analysis Plan (SAP): The written description of the statistical considerations for analyzing the data collected in the study. Includes how data are analyzed, what specific statistical methods are used for each analysis, and how adjustments are made for testing multiple variables. If some analysis methods require critical assumptions, the written description should allow data users to understand how those assumptions were verified.
- Informed Consent Form (ICF): The final version of the legal document approved by a human subjects protection review board. It is written in lay language and describes, among other things, the study's purpose, procedures, risks and potential benefits.
- Study Protocol with SAP and/or ICF: The study protocol that also includes a statistical analysis plan (SAP) and/or an informed consent form (ICF). Select one or both.
    - Statistical Analysis Plan (SAP)
    - Informed Consent Form (ICF)

Background: In a variety of visual search experiments, both basic and clinical, the data have been consistent with a situation where the variability of the signal (or target) is greater than the variability of the noise (distractors). As an example, in screening for breast cancer, this would mean that the variability of backgrounds in mammograms would be less than the variability in actual signs of cancer (the signal in these "signal detection theory" (SDT) terms. The classic sign of this greater signal variability is a "zROC" function with a slope < 1 – typically around 0.6. A slope of 1.0 is indicative of an equal variance 2AFC task; that is, the same amount of variability in the signal and the noise. The present experiment is part of a series where we systematically the parameters of a simple 2-alternative forced-choice (2AFC) task where participants decide if a stimulus is "good" or "bad". Participants can also be helped by a simulated AI system. This our **Human-AI Collaboration Tester (HAICT).**

For the HAICT task that we have been testing in this series of experiments, we would expect equal variance, but we think it would be worth checking so we will systematically vary prevalence which will shift criterion. That will sweep out an ROC curve that we can examine.

We will also test the Second Reader faux-AI in order to determine if low prevalence makes Second Reader worse.

- (H1): We expect to replicate the finding that human criteria become more conservative as prevalence declines.
- (H2): We predict that the slope of the resulting zROC will be 1.0.
- (H3): We hypothesize that low prevalence will make Second Reader AI less effective because the positive predictive value of its comments will be low.

1. **Dependent variable.** Describe the key dependent variable(s) specifying how they will be measured.

The main dependent variables of interest are accuracy (and the signal detection derivatives of accuracy, d' and c), reaction time, and subjective ratings on the survey following each block.

2. **Conditions.** How many and which conditions will participants be assigned to?

This series of experiments investigates how changing the input from a simulated AI can affect the decisions made by human observers in a two-alternative forced choice task (like the decision to recall a woman for further examination in mammography). We have developed a paradigm called the Human-AI Collaboration Tester (HAICT) that allows for efficient testing of interactions between a human and a simulated AI.

The observers' task in all conditions is to give a 2AFC decision about whether a stimulus is "bad" or "not bad." To use language roughly mimicking a medical diagnosis, each stimulus is referred to as a "case." Observers are asked to make a 2AFC decision about arrays of colored shapes. The decision is made based on the predominant color of the case. The number of elements of each color are drawn from one of two normal distributions, one for positive (bad) stimuli and the other for negative (not bad) stimuli.

The results from previous HAICT experiments (3 and 4) showed that human performance in the Second Reader condition drops off significantly at low prevalence. Performance in the Second Reader condition was better than Baseline when the prevalence of bad cases was 50% but was significantly worse than Baseline when prevalence was only 10%. In this experiment, we manipulate the prevalence of "bad" cases in the Second Reader and Baseline conditions. Four different prevalence rates will be tested – 10%, 33%, 67%, and 90%. Observers will complete 8 blocks (2 AI rules x 4 prevalence rates), and block order is random.

AI rules to be tested:

1. Baseline - No AI input. Observer classifies each case as "bad" or "not" bad on their own.
2. Second Reader - The observer makes an initial decision about every case. The AI silently classifies stimuli using a conservative criterion (c = 0.5). The logic for the conservative criterion is that the second reader is being used to cut down on false positive responses and so it is intended to question positive human responses that

might be marginal. If the observer and AI disagree, then the AI informs the human observer. The observer is then given a chance to either change their response or go with their first opinion.

As in Experiments 1-5, the AI d-prime is fixed at 2.2. Feedback is known to increase the prevalence effect, so feedback will be given in both the practice and the test trials. Observers will complete 20 practice trials and 200 test trials in each block. Immediately after each block is completed, observers will be shown a summary of their performance. After the Second Reader blocks, they will also be asked to answer three subjective questions about the usefulness of the AI (see "Files" for more details).

3. **Analyses.** Specify exactly which analyses you will conduct to examine the main question/hypothesis.

First, we summarize the number of hits, true negatives, misses, and false alarms in each block. From this, we can calculate the accuracy, the positive predictive value, sensitivity (d-prime), and the criterion for each observer under each of the different conditions. Given measures of performance at 4 levels of prevalence, we can estimate the ROC curve (pHit x pFA) and the zROC function (zHit x zFA). We will test the hypothesis that the slope of the zROC is equal to 1 (the consequence of an equal variance 2AFC task).

4. **More analyses.** Any secondary analyses?

We will look to see if the observers' subjective opinions about the AI are correlated with variables such as the empirical d-prime, or the positive predictive value.