Department of Transportation
Federal Railroad Administration
1**INFORMATION COLLECTION SUPPORTING STATEMENT B**
Experimental Investigation of Automation-induced Human Error in the Locomotive Cab
OMB CONTROL NUMBER 2130-XXXX

1. Description of sampling method to be used.

The participants will be identified from a pool of current engineers and conductors. Volpe personnel have experience in contacting the railroad personnel regarding a voluntary study, and this communication (email) will be sent to a local (to the Boston area) freight railroad. The same mail will be used to recruit engineers and conductors from the Fort Worth, Texas area. These two sets of operators were chosen for their experience on the Trip Optimizer and Electronic Train Management System (ETMS) Positive Train Control (PTC), the types of automation installed in the Cab Technology Integration Laboratory (CTIL) simulator. Participants will need to be professional engineers or conductors who have at least 1 year of experience working with the TO or PTC technology.

2. Description of procedures for information collection, including statistical methodology for stratification and sample selection.

The technology we will be investigating in the research (PTC-ETMS and TO automation) currently exists and is implemented in locomotive cabs across the country. These are at a Technology Readiness Level (TRL) of 9. The proposed project is for human in the loop research. The technology we will develop will be in the form of HAI design guidelines and proposed design changes that could potentially be mocked up and evaluated in a subsequent experiment. These potential results would be at a TRL of 4, a static system/subsystem model or prototype demonstration in a relevant (but not operational) environment, e.g., the CITL.

TO has been developed for the purpose of saving fuel, and railroads have been implementing the technology in their locomotive cabs. However, engineers have expressed concerns with the increased monitoring requirements. Based on the errors observed in a brief pilot study, it appears that these concerns warrant further investigation to understand their causes in automation design and in the operating environment. The proposed research is an in-depth and more carefully controlled experimental study that extends that initial work.

PTC, in contrast, is designed to enhance safety and prevent overspeed, incursions into workzones, and collisions. The differing goals of two technologies, productivity (TO) versus safety (PTC), has sometimes been found, in the transportation industry, to compromise safety. It is also fully expected that an eventual integration of PTC and TO would have far more than two modes of operation. This situation – automation with potentially conflicting goals and multiple modes of operation – can easily lead to human error and requires investigation

This proposed research is a systematic empirical HITL study to be conducted in the Volpe CTIL simulator, with twenty-four 2-person crews of operators (48 total participants). This sample size was chosen based on the experimental design and the anticipated number of subjects to provide the appropriate statistical power to draw meaningful conclusions from the data. See below for additional explanation. We will use a similar scenario to the one described in the FRA's previous experiment (Sebok, Walters & Wickens, 2017). The scenario will include a 17-mile segment of

track with a divergence in the track. At some point during the scenario, TO will request that the engineer provide track information, and PTC will make a similar request for engineer input. If the engineer fails to input that information within a limited amount of time, the automation requests the information in a more salient manner (in accordance with current implementation practices), and then – if the engineer still does not provide input – TO initiates a switch to manual mode. We will include a carefully timed distraction just prior to and co-incident with the track request in the high workload condition.

The hypotheses we will investigate are:

H1a) Automation provides specific performance benefits (e.g., TO reduces fuel usage; PTC prevents overspeeding and transgressions into workzones or past a red signal) compared with manual control.

H1a will be evaluated by comparing performance in the manual versus automated conditions. Performance will be assessed using multiple measures: number of safety violations (e.g., incursion into a workzone, passing a red signal, passing a stop and protect crossing, overspeeding), fuel usage (for trip optimizer), and train handling characteristics (e.g., throttle / braking cycles, where longer and fewer cycles indicate better control of the train; forces on the couplings between cars).

H1b) Automation does not reduce perceived workload in the locomotive cab compared with manual control.

This will be evaluated by comparing subjective workload (assessed using the NASA TLX inventory) after each scenario. TLX scores will be compared across the automated high and low task loading conditions, and the manual high and low task loading conditions.

H2) Automation condition will show more errors in high workload situations than in low workload situations (e.g., distractions lead to failure to notice mode transitions) and these manual condition will not.

	Manual	***Automated***
Low Workload	24 crews	12 crews PTC, 12 crews TO
High Workload	24 crews	12 crews PTC, 12 crews TO

We anticipate that the issue of perceived workload is more nuanced than the hypothesis currently states. For example, automated systems reduce aspects of workload. TO reduces the need for the engineer to adjust the throttle and braking but increases need for systems monitoring and verification. PTC does not affect the tasks associated with train handling, but it might reduce the perceived workload associated with monitoring for speed restrictions and red signals while increasing the perceived workload associated with monitoring systems.

Note that the first hypothesis, that automation does not reduce workload, contradicts much of the empirical research and operational experience with automation in aviation and other industries (e.g., Ferris, Sarter & Wickens, 2010). In general, the use of automation is associated with lower workload in decision making and control (but not visual monitoring) as the operator has fewer tasks to perform, as long as the automation is working appropriately. Concerns that are typically associated with the use of automation are that the operator can be out of the loop and lose awareness of what is happening, and that the operator can suffer from manual skill degradation (Sebok & Wickens, 2017). The fact that the TO requires the engineer to assume a highly interactive supervisory role over the automation and take over the task of monitoring the automation and actively verifying that the recommendations are valid, imposes additional workload on the engineer. The effects of PTC are less clear: as a safety system, it provides an additional protection, but it also offers another system to be monitored. It does not appear to – in routine circumstances – take over tasks performed by the engineer. However, the industry has not had the same reaction to PTC automation.

The data in support of the second hypothesis (H2) come from the studies of aircraft automation by Sarter et al., 2007 and Dehais et al., 2017, described previously.

Importantly, research on human-automation interaction errors has identified 3 important classes of errors, all of which can be invited by locomotive automation:

1. Set up errors, whereby the human programs the automation to do something in a way unintended (e.g., enters the wrong mode or parameters).

2. Complacency errors, whereby the human fails to monitor the automation with sufficient vigilance, and hence fails to detect a condition that the automation is unable to function appropriately or fails to notice that the automation has itself changed modes, without human intervention.

3. Mode errors, when the operator thinks the automation is in one mode (e.g., PTC on) when it is in fact in another mode (e.g., PTC off), and responds appropriately for the thought-of mode, rather than the actual mode. Clearly, mode errors can often be a consequence of complacency errors.

In our simulation, we will design and introduce conditions that induce each of the three error types and investigate if and how they occur in both low and high workload scenarios. The proposed study will include 4 conditions: two manual conditions in which no automation is present, and two automated conditions (TO or PTC, depending on the crew's expertise). There will be a low workload manual condition and a high workload manual condition. There will be a low workload automated condition and a high workload automated condition.

Table 1. Experimental Design

	Manual	**Automated**
Low Workload	24 crews	12 crews PTC, 12 crews TO
High Workload	24 crews	12 crews PTC, 12 crews TO

The figure below shows the scenario events for the Sebok, Walters & Wickens, 2017 study. The top part of the figure, above the blue dashed line, shows events in the "Low Workload" condition and the bottom part of the figure, below the blue dashed line, shows the events in the "High Workload" condition. The currently proposed study will use a similar event plan.

The Low Workload conditions will include multiple speed restrictions and a quiet zone. In the Manual condition, there will be no prompts from TO or PTC automation, as these systems will not be operating in the Manual condition. In the Low Workload automated condition, there will be prompts from automation if appropriate for the scenario (e.g., a TO "prompt for track information" in the case of an upcoming track divergence).

The high workload condition will include all of the low-workload events (speed restrictions and quiet zones), but it will also include additional communications and events (e.g., workzone, restricted speed zone, and gate crossing failure).

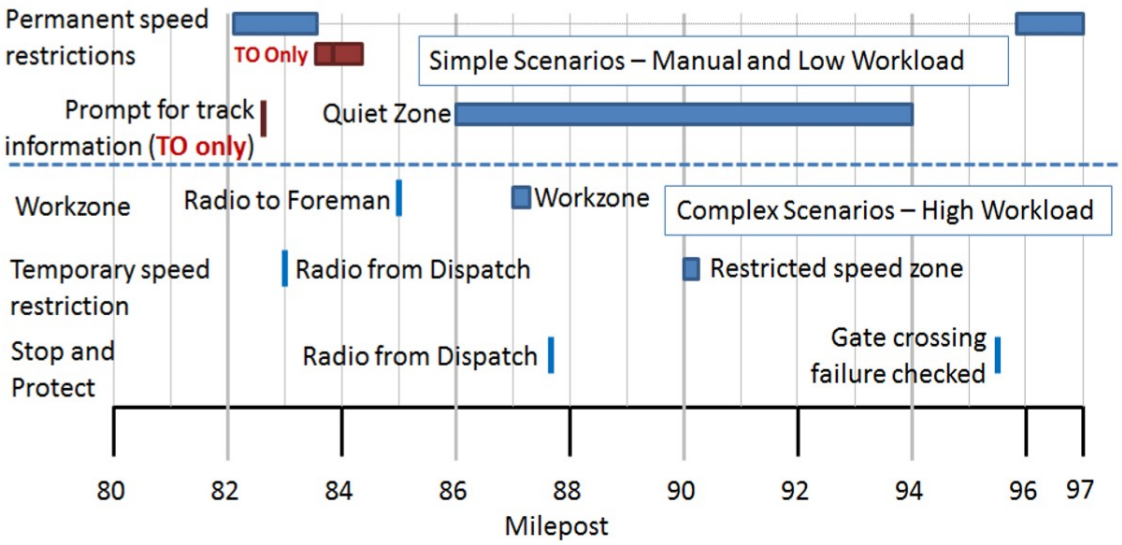

Figure 2: Events by trip milepost for the simple and complex scenarios.

The distractions in the currently proposed study will be identical in nature to the distractions used in the Sebok, Walters & Wickens (2017) study. Just before the locomotive cab automation presents an important visual change, a call from a dispatcher (simulator instructor research confederate) will be made to the locomotive cab. This will be a radio (audio) call that both the engineer and conductor will hear.

The *difference* between this proposed study and the previous study is that, in the proposed study, the dispatcher communication will *be carefully timed* to occur just prior to the automation change. In the previous exploratory study, the communications were not tied to any events in the automation.

We will use the standard $p < 0.05$ for significance, and $p < 0.10$ to indicate a trend. The number of errors (tallied for each condition: manual low workload, manual high workload, TO low workload, TO high workload, PTC low workload, PTC high workload) will be compared via t-tests.

In the pilot evaluation, we did not evaluate effect sizes. It was simply an opportunity to investigate human performance with the locomotive cab automation and we happened to observe errors that require further investigation. The scenarios in the pilot evaluation were designed to include the same events (e.g., all high-workload scenarios had the same speed restrictions and dispatcher-initiated tasks) but the timing of these events was not controlled precisely, as will happen in the proposed experimental study.

In one scenario in the pilot evaluation, the dispatcher radio call happened to occur at the same time as an automation request appeared on the TO display. The engineer, distracted by the radio call, missed the request on the TO display. We are implementing that carefully timed distraction into the design of the proposed experimental study. The fact that we found errors in the investigative analysis (where we did not implement tight experimental control) suggest that we should expect to see errors in a carefully designed study with subject matter expert (SME) input. To validate the workload manipulations, we had a SME perform the scenarios and indicate his impression of the workload. We will perform pilot testing to provide another validation, and we will collect subjective workload data during the proposed experiment.

No statistical analyses were performed in the pilot study. It was a qualitative investigation.

For the analyses in the proposed experiment:

The primary errors we are looking for in the scenarios have a limited number of opportunities to appear. We will use repeated measures t-tests to compare performance across the 4 different conditions.

1) Failure to notice automation set-up and mode transitions.
    a. 2 setup errors per crew
    b. 2 automation mode transitions per crew
2) Failure to notice trespassers / gates up.
    a. 8 opportunities total, per crew. (2 per scenario in each of 4 scenarios)
3) Failure to make necessary communications with dispatch and work crew foremen.
    a. 4 requirements per crew (related to contacting work crew / dispatch)
    b. 8 requirements regarding trespasser and gates
4) Failure to stop appropriately at a stop and protect.
    a. 2 opportunities per crew

For train control parameters, we will use qualitative expert assessment to characterize the performance differences across the conditions. We do not anticipate performing statistical analyses on these results. It is possible that we will develop a 5-point rating scale to assign to "train control handing quality," and we will use t-tests to compare performance across conditions.

Other errors that might occur during the scenario include overspeed, incursion into a work zone, or failure to consider the implications of an automation mode change and take appropriate action. These will simply be reported, as they are expected to be highly infrequent (e.g., perhaps 2-3 total in the experiment). If there are more errors noted, we will tally and use t-tests, as described above for the other error types.

- We used lessons learned from the pilot study and worked directly with a SME to set up the experimental manipulations. Therefore, we expect that our manipulations will have at least medium to large effects, as measured by Cohen's D (0.5-0.8). The assumption of a medium effect is probably conservative, given the SME input in the scenario design.

- We used G*Power, a common power analysis tool, to perform a post-hoc power analysis for our repeated measures t-test design. See Figure 1 for the input and output parameters of the power analysis. Given a sample size of 24, and an assumed medium effect size of 0.6, we obtain Power of 0.88, which is well within the accepted 0.8-0.9 range for desired power.

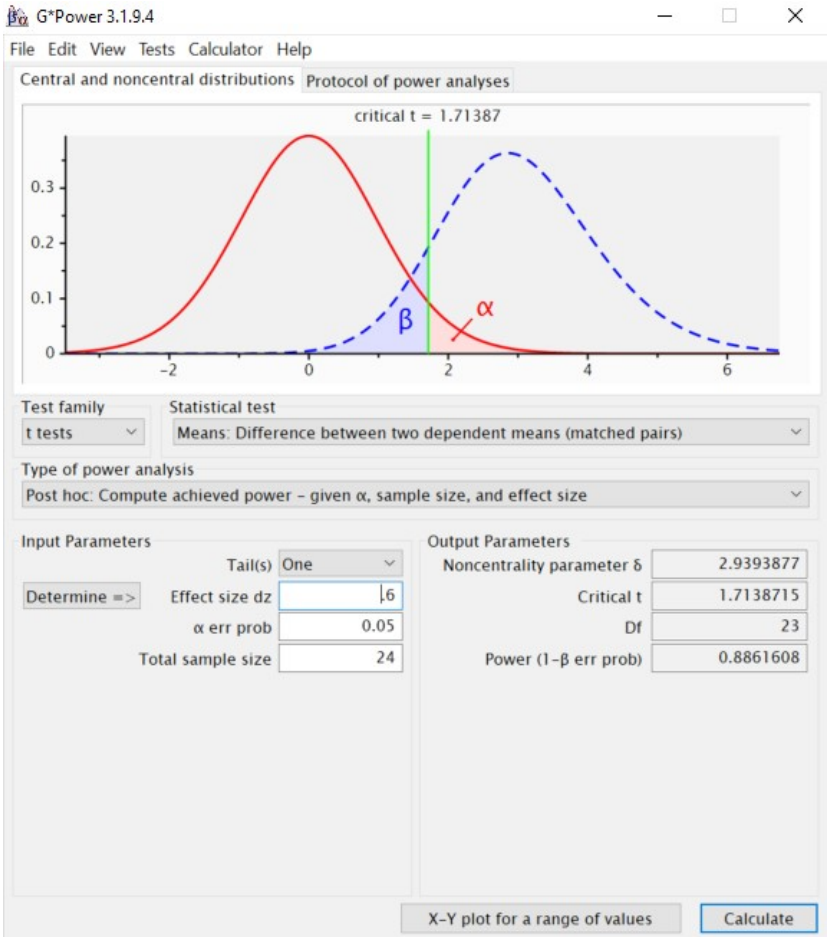

Figure 1: Screenshot from G-Power Software.

3. <u>Description of methods to maximize response rate and to deal with non-response issues</u>.

As this research is not survey research, response rate and consideration for non-response is not an issue. Given response to this question, all volunteers for the study will be monetarily compensated for their participation

4. <u>Describe any test procedures for procedures or methods to be undertaken</u>.

The study will include 5 sessions in which participants (2-person crews) run a simulated train along a 17-mile segment of track. The participants will receive information about the study, and be asked to sign consent forms. They will perform a practice (familiarization) run to get a feel for the consist dynamics and the grade. They will then perform 4 scenarios on that same section of track, with the same simulated consist. The scenarios will include (in counterbalanced order) two manual mode and two automated mode (either PTC or TO, depending on the crew's expertise). Following each scenario, participants will complete the NASA-TLX workload inventory. Participants will be given a 5-minute break after each scenario. After completing the final scenario, the participants will be debriefed about the study, and the researchers and crews will discuss the scenarios and the crew's operational experiences with automation in the locomotive cab.

FRA will use the NASA TLX subjective workload inventory following each scenario to assess each ***participant's perceived workload*** in that scenario. We will average across the 6 scales (inverting "performance" – the only scale where "higher means better") and compare the averages across the 6 conditions. This will identify if the participants experienced different degrees of workload in the conditions, and it will provide a "check" on the proposed experimental manipulation.

Another performance metric will be subjective evaluation of train control parameters. These will be plotted and evaluated in collaboration with rail subject matter experts to identify qualitative differences in train handling across the conditions. Depending on the results, we may use quantitative metrics (to be defined) such as oscillations or fluctuations in handling characteristics (e.g., throttle / braking cycles, where longer and fewer cycles indicate better control of the train) or the amount of fuel consumed (as determined by the simulator), or overspeed amounts, or degree of incursion into a workzone.

5. <u>Provide name and phone number of individuals consulted on statistical aspects of study design and other persons who will collect/analyze information for agency</u>.

None consulted beyond original research proposer. The proposer for this research, Tier 1 Performance Solutions will do all collection and analysis of data.
Study lead at Tier 1 is Angie Sebok, 1-720-699-1509

Point of contact for the study:

Michael E. Jones
US Department of Transportation
Federal Railroad Administration
Human Factors Division (RPD-34)
Washington, DC 20594
Michael.e.jones@dot.gov
202-493-6106