Bridging the Simulator Gap: Measuring Motivational Bias in Digital Nuclear Power Plant Environments

Rachel Benish Shirley, The Ohio State University[1]
Dr. Carol Smidts, The Ohio State University

Key words: *Nuclear Power, Human Reliability Analysis, Simulator Studies, Digital Simulator Environments, Structural Equation Modeling, Latent Variables, Mediation, Sample Size*

ABSTRACT

Using digital NPP simulator facilities for research in Human Reliability Analysis (HRA) and nuclear power research is increasingly popular. We propose a method for characterizing and quantifying the gap between data collected in a simulator and data that reflect NPP operations. Using novice operators, we demonstrate how to manipulate and measure the impact of the simulator environment on operator actions.

A set of biases are proposed to characterize factors introduced by the simulator environment. There are two categories of simulator bias: **Environmental Biases** (physical differences between the simulator and the control room), and **Motivational Biases** (cognitive differences between training in a simulator and operating a NPP). This study examines Motivational Bias.

A preliminary causal model of Motivational Biases is introduced and tested in a demonstration experiment using 30 student operators. Data from 41 simulator sessions are analyzed. Data include crew characteristics, operator surveys, and time to recognize and diagnose the accident in the scenario.

Quantitative models of the Motivational Biases using Structural Equation Modeling (SEM) are proposed. With these models, we estimate how the effects of the scenario conditions are mediated by simulator bias, and we demonstrate how to quantify the strength of these effects.

1 INTRODUCTION

The proposed research is based on the premise that, despite fifty years of research and development, Human Reliability Analysis (HRA) continues to be the weak link in Probabilistic Risk Assessment (PRA) for commercial NPPs in the United States and around the world.

[1] Corresponding author. Email address: rachel.e.b.shirley@gmail.com

NPP simulators present a potential solution to this problem. Recently, the Nuclear Regulatory Commission (NRC) developed the Scenario Authoring, Characterization and Debriefing Application (SACADA) "to address this data need" [1] using NPP training simulators.

One aspect of simulator data that is not accounted for by SACADA's data collection tool is the bias in simulator data due to the fact that data are collected in a simulator, not an operational NPP control room. The inherent drawback of simulator data is that the data is affected by the artificial nature of the simulator environment. Simulator bias is widely acknowledged but has not been systematically addressed. This paper introduces two tools to begin to address this gap in the research. The first tool is a framework for discussing the aspects of simulator bias that were first introduced in [2]. Two sets of biases are defined: Environmental Biases (effects of physical differences between the control room and the simulator) and Motivational Biases (effects of working in an artificial setting). This framework is provided in Section 1. The second tool is a method for measuring, modeling and quantifying the effects of the simulator environment on data collected in the simulator. We do this through a proof-of-concept experiment conducted using student operators in the Ohio State University (OSU) NPP Simulator Facility. The bulk of the paper focuses on designing the experiment and collecting data (Section 2), using Structural Equation Modeling to determine the model structure and estimate model coefficients (Section 3), and analyzing the model results (Section 4).

This experiment uses student operators and is not meant to provide reliable data on bias in studies of experienced operators; however, the approach we introduce and the analysis of bias in the student operator studies is offered as a guide for future studies that are valid in a broader setting.

1.1 Inherent Bias in Simulator Studies: Environmental and Motivational Biases

Although simulator study practitioners acknowledge the bias introduced by simulator studies (for example, [3]), recent literature has little to say about the necessary limitations of simulator studies. We address this gap by introducing a set of simulator biases taken from a literature review of flight simulators, driving simulators, and anesthesia simulators. The set of biases presented in Table 1 is provided as a framework for formal thinking about simulator bias. We hope that fellow researchers will add new biases and refine the definitions in Table 1 as this body of research grows.

We have identified two broad categories of biases that are necessarily present in a simulator study [4]. **Motivational biases** refer to the factors that motivate an operator's actions in a simulator that differ from the motivating factors in the actual plant. **Environmental biases** refer to any physical or environmental differences between operating the actual plant and operating the simulator. Table 1 summarizes the biases identified in [2] as likely to be present in NPP simulator studies [4].

Table 1: Motivational and Environmental Simulator Biases

Environmental Biases		**Motivational Biases**	
Context Effects	Any discrepancy between the control room and the simulator	*Hypervigilance*	Tendency to be vigilant or fastidious due to being observed or expecting an adverse event
Technology bias	Impacts from the technology used to simulate the simulator environment	*Cavalier behavior*	Tendency to be lax or incautious because the simulation is not real
Simulator sickness	Physical effects introduced by the simulator environment (headache, motion sickness, fatigue, etc.)	*Incentive effects*	Artificial boost in vigilance due to incentive

Environmental Biases		Motivational Biases	
		Policy response bias	Attempt to meet researcher's (stated or perceived) expectations
		Prominent hypothesis bias	Expectation of a certain accident or condition leads subjects to ignore certain indicators and pay extra attention to others.

1.1.1 Motivational Biases

Motivational biases are internal effects that are the result of the artificial nature of a simulator exercise. The Motivational Biases include Hypervigilance, Cavalier Behavior, Incentive Effects, Prominent Hypothesis Bias, and Policy Response Bias.

1.1.1.1 Hypervigilance & Cavalier Behavior

Hypervigilance refers to research subjects' tendency to be vigilant and fastidious because they know they are being observed or because they expect an adverse event to occur. NPP operators who are being tested in a simulator are expecting to be challenged, whereas a normal day at the plant will not provoke anticipation of adverse events.

The opposite of hypervigilance is cavalier behavior, that is, an operators' tendency to be cavalier because they know a bad outcome in the simulator has no real-world consequences. Performance-based incentives can counteract cavalier behavior, but these incentives must be carefully balanced to ensure they do not (re-)introduce hypervigilance [5].

In one study for assessing the authenticity of human simulation studies in anesthesiology, researchers found that a recording camera was a significant source of distraction in 26% of the scenarios. Researchers also observed that in at least 64% of the scenarios reviewed, the anesthesia provider anticipated that something would go wrong in the case. Researchers also noted "overattention" to monitoring values in nearly 70% of the case scenarios [5].

A typical mitigation strategy for hypervigilance is to introduce a distracter task to help the practitioner become more absorbed in the work and less aware of the simulator context. In the anesthesia study, the anesthesiologist was asked to supervise a junior colleague, a task that provided a realistic distraction and reduced the hypervigilance.

In a nuclear power plant setting, a simple but effective distracter task is requiring operators to perform a plant maintenance procedure. The procedure must be carefully selected to ensure that the maintenance operation will not influence the plant state during the period of interest unless the transition from the maintenance task to the accident response is specifically an object of study.

1.1.1.2 Incentive Effects

The incentives introduced to combat cavalier behavior may introduce a further bias, inducing subjects towards hypervigilance or otherwise artificially improving performance. The most effective mitigation strategy is to eliminate incentives; when this is not possible, the approach is to introduce incentives that are sufficient to combat cavalier behavior but not strong enough to introduce hypervigilance. This is a difficult structure to balance, but separating simulation results from the incentive seems to be an effective strategy. Subjects understand that they will receive the incentive if they give an earnest effort without needing to be "perfect."

1.1.1.3 Prominent Hypothesis Bias

Another factor related to hypervigilance is the Prominent Hypothesis bias. Operators may anticipate a specific accident or failure, leading them to look for indicators of this accident and even to ignore contrary indicators [6]. Prominent hypothesis bias is similar to the cognitive bias known as anchoring; when solving a problem, people start with an initial "anchoring" solution in mind and adjust from that anchor rather than evaluating an objective situation [7]. While anchoring bias can occur in the real world (for example, operators that have had trouble with the steam generators might be more prone to assume an accident is an SGTR), the prominent hypothesis bias refers to an anchor introduced by the simulator environment. The prominent hypothesis bias is a double-edged sword: When the hypothesis is correct, a strong Prominent Hypothesis will improve the subjects' performance. In a NPP simulator, the operators might identify the accident quicker and proceed immediately to the appropriate corrective action. However, operators who are focused on an incorrect hypothesis must overcome their initial expectations in order to recognize the accident and take the appropriate steps.

1.1.1.4 Policy Response Bias

Also tied to hypervigilance, policy response bias reflects an operators' tendency to perform according to policy rather than to operate as he or she would while actually operating a plant [8]. Policy Response Bias is also referred to as Demand Characteristics in psychology research [9]. For example, an operator being observed by an NRC evaluator is likely to follow procedures and not skip procedural steps, even if the operator might skip those steps while actually operating the NPP. One mitigation strategy is to make studies truly anonymous. This factor is more relevant when observing professional operators in a professional context; student operators are not likely to have the same concerns.

1.1.2 Environmental Biases

Environmental Biases are the effects of gaps or variations between the control room and the simulator control room. The Environmental Biases include Context Effects, Technology Bias, and Simulator Sickness.

1.1.2.1 Context Effects

Context Effects are the effects of any discrepancies between the actual simulator and the plant control room [10]. In high fidelity simulators used in commercial plants, an example of context effects is any discrepancy between the plant reaction time and the plant responses programmed into the simulation software. This impact is seen in the OSU simulator facility in the limited display space available. Although most of the control panels are displayed on the large, wall-mounted screens, these displays are too small to read alarm labels or discern plant parameter values. Operators are forced to scroll through the workstation displays to identify actual values. This can be mitigated by adding additional displays. In general, context effects are mitigated by building the simulator to match the actual control room in as many ways as is possible. An alternative approach is to recruit operators who are familiar with operating a variety of plants and working in a variety of settings; the more familiar an operator is with adapting to new control room environments, the easier it will be for the operator to perform at high levels in the simulator.

1.1.2.2 Technology Bias

Technology Bias is a subset of Context Effect. Working on a digital platform that simulates an analog environment introduces an inherent technology bias: operators may be better-able to operate the analog plant than its digital counterpart, or vice versa [11]. This bias is not present in high fidelity

simulators that use the same technology and control interface in the simulator as in the control room. However, digital simulators such as the OSU NPP Simulator Facility are digital facsimiles of analog controls. The new technology used to interact with the plant introduces artificial effects. For example, in the OSU NPP Simulator Facility, student operators had trouble activating some switches or understanding how to open or close a valve. These actions would be more intuitive in the physical plant, where a tactile response provides feedback that is absent in the digital simulator.

Technology bias becomes important when working with experienced operators who have trained and worked on analog plants. In such a study, the computer interface may significantly impact study outcomes. To reduce this bias, even experienced operators who are familiar with the plant layout must be trained on the digital simulator platform before a study is conducted.

1.1.2.3 Simulator Sickness

Simulator sickness is typically associated with flight simulators and other motion-based simulators, where practitioners may be subject to nausea and motion sickness induced by the moving displays [12]. While motion-related sickness is not anticipated in OSU's NPP Simulator Facility, a few physical side effects have been observed. A few student operators have complained of headaches induced by the computer screens. The simulator environment should be evaluated to determine if air quality, noise levels, etc. will impact the research subjects.

2 METHODS AND MATERIALS

Using the list of motivational biases in Table 1, we manipulate simulator conditions to test our assumptions of how the Motivational Biases impact operators in a simulator. We do this by observing student operators responding to accident conditions in the OSU NPP Simulator Facility.

We introduce experimental bias manipulations and construct a causal model of their expected effects. We then observe student operators in the simulator under each bias manipulation. After each scenario, we survey the student operators to quantitatively assess the simulator bias. We also measure two response times as a proxy for operator performance. From this data, we estimate the parameters for two Structural Equation Models (SEMs) of simulator bias: the Latent Bias Model and the Bias SEM.

Section 2.1 describes the bias manipulations, the experimental conditions, and the data used to construct the bias models. Section 2.2 provides a brief overview of SEM and defines the two models developed in this experiment.

These models are based on student operators and are not expected to represent experienced operator behavior. However, the results in Section 3 and the analysis in Section 4 are included as an example of how this approach can be used to assess and understand the impact of simulator bias on data collected in NPP simulator facilities.

2.1 EXPERIMENT DESIGN

The bias manipulations introduced in this experiment are labeled X1 – X4 in the diagrams throughout this paper:

- *X1, Trt1*—the mix of scenarios the operators participate in (half the operators participate in a mix of familiar and unfamiliar scenarios, while the other half participate only in familiar scenarios)
- *X2, Graded*—whether the data are collected in practice sessions or during the final exam
- *X3, Fami*—the operator's familiarity with the accident scenario.
- *X4, Yr2015*—the operator's training, determined by the year the operator participated in the OSU NPP Systems and Operations Course. The 2015 course was more comprehensive and included additional time to practice and train in the simulator. *Yr2015* can be understood as an indicator of how additional experience will impact the bias in the scenario.

These manipulations are described in more detail in Table 2.

Table 2: Variables in the OSU NPP Simulator Experiments - Experimental Manipulations

Variable Name	Description
Experimental Manipulations *(Controlled for in the experimental design)*	
[X1] Trt1	Crews are assigned to Treatment 1 or Treatment 0. Treatment 0 crews respond to scenarios with *Fami* = 0 or 0.25; Treatment 1 crews encounter a mix of familiar and unfamiliar scenarios (*Fami* = 0.0, 0.25, 0.5, 0.75 or 1.0). In the SEM, *Trt1* = 1 for Treatment 1 crews and 0 for Treatment 0 crews.
[X2] Graded	Data are collected during two phases of the experiment: Practice sessions and final Exam sessions. In the SEM, *Graded* = 1 for Exam sessions and 0 for Practice Sessions
[X3] Fami	Familiarity of scenarios is rated on a scale from 1 (least familiar) to 5 (most familiar). Familiar scenarios are scenarios that operators studied in the NPP Operations course; unfamiliar scenarios are complications of the more familiar scenarios. In the SEM, *Fami* is normalized on a 0 to 1 scale, with values of 0, 0.25, 0.5, 0.75 or 1.0.
[X4] Yr2015	Data are collected in 2014 (*Yr2015 = 0*) and 2015 (*Yr2015 = 1*). *Yr2015* is a measure of training; in 2015, students spent more time in the simulator and had more training related to evaluating the significance of an accident.

Figure 1 is a causal model that illustrates how the four manipulations are expected to impact the Motivational Biases in the OSU NPP Simulator Facility Experiment. Rectangles in Figure 1 represent Experiment Conditions (*Graded, Trt1, Fami*), ovals are the biases that are manipulated by the experimental conditions, and hexagons are the survey responses used to evaluate the biases. *Yr2015* is not included in the figure because this variable may impact all aspects of the bias.

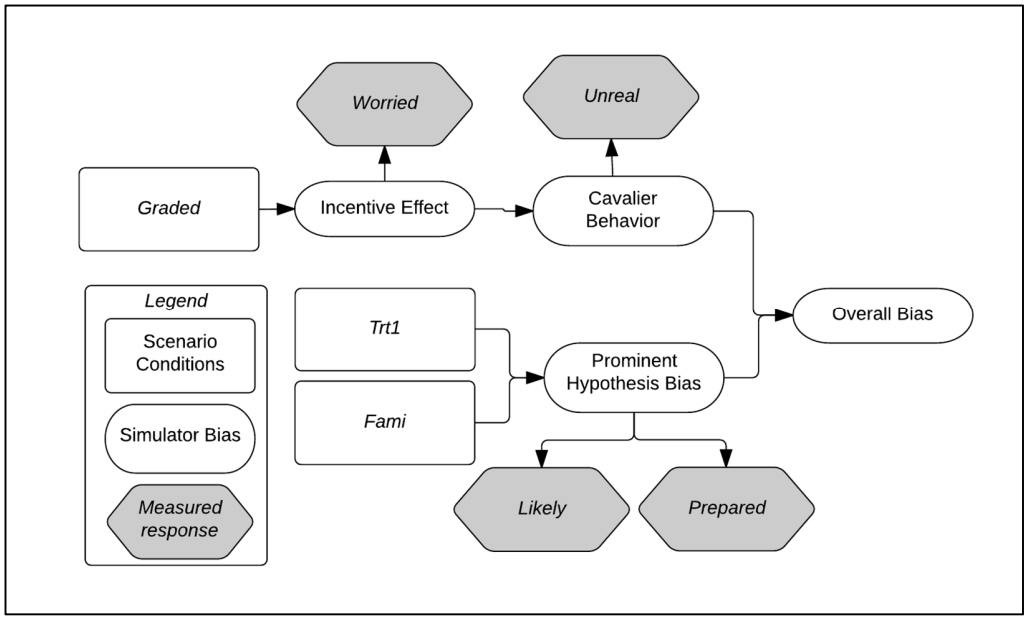


Figure 1: Causal model of simulator bias in the OSU NPP Simulator Experiments

The Motivational Biases that are manipulated in the OSU NPP Simulator Experiment are *Incentive Effects*, *Cavalier Behavior*, and *Prominent Hypothesis.* Simply being in the simulator environment triggers *Cavalier Behavior*. The *Graded* manipulation counteracts *Cavalier Behavior* by triggering the *Incentive Effect* when the students are taking their final exam.[2]

Fami and *Trt1* manipulate *Prominent Hypothesis*. When *Fami* is low, operator *Prominent Hypothesis* is expected to be low, and vice versa. Similarly, students in *Trt1=0* crews are expected to have a stronger *Prominent Hypothesis* than those in *Trt1=1*, who encounter new and unknown scenarios.

Table 3 summarizes the conditions introduced in Figure 1, their expected effects on the simulator bias, and the bias measures used to evaluate each bias.

Table 3: Hypotheses in the OSU NPP Simulator Experiment

Condition	Expected Effect	Measurement
Simulator environment	Cavalier behavior	*Unreal*
Phase (*Graded*): Practice or Exam	Incentive Effect, counteract cavalier behavior	*Worried*
Familiarity (*Fami*)	Prominent hypothesis bias decreases as familiarity decreases	*Likely*, *Prepared*
Treatment (*Trt1*): 0 – all familiar accidents, or 1—mix of familiar and unfamiliar	Prominent hypothesis bias lower in *Treatment* = 1	*Likely*, *Prepared*

In total, five bias measures are introduced to estimate the bias effects: *Unreal, Prepared, Likely.For, Likely.Against,* and *Worried*. The bias measures are based on survey responses provided by the

[2] Of course, incentives to reduce *Cavalier Behavior* can result in *Hypervigilance*. In this preliminary analysis, we do not attempt to differentiate between responsible operator behavior and *Hypervigilance*.

operators after the simulator session. Two bias measures—*Likely.For* and *Likely.Against*—are derived from the *Likely* survey question. These and the other three bias measures are defined in Table 4.

Table 4: Variables in the OSU NPP Simulator Experiment - Bias Measures

Variable Name	Description
Bias Measures *(Operator-Reported Values, averaged between the Senior Reactor Operator and the Reactor Operator)*	
[M1] Unreal	A measure of overall bias, *Unreal* is the operator's perceived reality of the scenario. Survey question: "During the scenario, how real did the simulator feel?" (scaled response; 0 = I felt like a college student in a simulator, 10 = I felt like an operator in a nuclear power plant). We *Unreal,* i.e. 10-*Real,* rather than *Real* because *Bias* is expected to increase when the scenario feels more artificial. The data are sorted into 5 ordinal levels: $\boldsymbol{UnReal}: Real \leftarrow 1-2-3-4-5 \rightarrow Not\ Real$
[M2] Prepared	*A measure of prominent hypothesis bias. Survey question: "How prepared were you to participate in this scenario?" (scaled response; 0 = "Completely unprepared. I would need significant additional training to feel prepared in this scenario," 10 = "I felt completely prepared to respond to this scenario." The data are sorted into 6 ordinal levels:* $\boldsymbol{Prepared}: Not\ Prepared \leftarrow 1-2-3-4-5-6 \rightarrow Well\ Prepared$
[M3] Likely.For ***[M4] Likely.Against***	A measure of the Prominent Hypothesis Bias. Survey question: "With what likelihood do you expect to encounter the following accident conditions?" (rate each option: SGTR/SLB/LOCA/Other). This response is split into two dimensions: *Likely.For* is the bias that increases the operator's expectation of the event that occurs in the scenario, and *Likely.Against* is the bias that increases the operator's expectation of events that do not occur in the scenario. A nominal, unbiased response to the survey question would be [25% SGTR, 25% SLB, 25% LOCA, 25% Other]. Any change in this distribution that corresponds to the simulator scenario is counted as *Likely.For* bias, and any change that corresponds to a different scenario is measured as *Likely.Against.* For example, if the scenario is an SGTR and the operator expects [50% SGTR, 50% SLB, 0% LOCA, 0% Other], the operator is biased as follows:

Possible Event	***Unbiased Expectation***	***Perfect Bias (Event = SGTR)***	***Operator Expectation***	***Bias***
SGTR	0.25	1.0	0.50	0.25 *For*
SLB	0.25	0.0	0.50	0.25 *Against*
LOCA	0.25	0.0	0.0	0.25 *For*
Other	0.25	0.0	0.0	0.25 *For*
			Total	*1.0*
			For	0.75 *For*
			Against	0.25 *Against*

[M5] Worried	The maximum total bias is 1.5. Maximum total *Likely* bias occurs when operators report 100% expectation of one accident (e.g. [100% SGTR, 0% SLB, 0% LOCA, 0% Other]). We rescale total bias from [0,1.5] to [0,1] by dividing all bias estimates by 1.5. In the example above, this results in *Likely.For = 0.50* and *Likely.Against = 0.17*. *Likely.For* and *Likely.Against* are continuous variables restricted to a [0,1] scale.
	A measure of the incentive effect. Survey question: "Do you worry that your grades will be affected negatively if you don't perform well during the accident response?" (Yes or No). In the SEM, *Worried* = 1 if Yes and 0 if No. The data are averaged, resulting in a three ordinal levels: $\boldsymbol{Worried}: Not\ Worried \leftarrow 1-2-3 \rightarrow Worried$

Cavalier Behavior is measured by *Unreal*, the operators' perceived reality of the scenario. In future work, measures to target Cavalier Behavior specifically can be developed by modifying measures of safety culture and employee attitudes (for example, [13]).

Two Motivational Biases are not included in Figure 1: *Hypervigilance* and *Policy Response Bias*. *Hypervigilance* is included implicitly as the opposite of *Cavalier Behavior.* For clarity, we only use one

variable (*Cavalier Behavior*) in this analysis. *Policy Response Bias* is expected to be consistent throughout the experiment; none of the manipulations (X1-X4) are designed to alter possible policy response bias.

In addition to the variations in simulator bias captured in the Bias Causal Model (Table 1), we are also interested in the effects of simulator bias on the overall scenario. In this demonstration experiment, we use operator response times to examine the effects of simulator bias. We report two response variables: *Trip* and *Dif. Trip* is the time to trip the reactor. *Dif* is the time between the reactor trip and when the operators exit the Emergency Operating Procedure (EOP) E0 to enter an accident-specific procedure. These two times roughly correspond to the time to recognize that a serious accident has occurred (*Trip*) and the time to secure the plant and diagnose the accident (*Dif*). The response times are defined in Table 5.

Table 5: Variables in the OSU NPP Simulator Experiments - Response Times

Variable Name	Description
Response Times *(Response times are divided by the mean response time for the same scenario to minimize effects of differences between scenarios)*	
[R1] Trip	*Trip* is the scenario normalized time to recognize accident**.** *Trip* can be thought of as the time required to recognize and respond to the accident, relative to other crews responding to the same accident. Each scenario's accident requires shutting down the reactor. Eventually, the plant's engineered features will trigger an automatic reactor trip, but some crews are able to recognize the problem and trip the reactor before it reaches this level. Time to trip is normalized for each scenario by dividing by the mean trip time for the scenario. Note that instead of dividing seconds to trip by the mean for the scenario, normalized times could also be obtained by dividing by the expected time required, as estimated by a dynamic HRA model such as the Information Decision Action Crew (IDAC) virtual operator program [14].
[R2] Dif	*Dif* is the scenario normalized time to diagnose the accident. *Dif* can be thought of the time taken before operators begin corrective actions, relative to other crews responding to the same accident. After the reactor trips, operators use Emergency Operating Procedure (EOP) E0 to secure the plant, ensure safety functions are operating correctly, and diagnose the accident. This variable measures the time between when the reactor trips and the time when operators exit E0 and enter another procedure that is specific to the accident that occurred. In this research, *Dif* is the time until operators settle on a final diagnosis, marked by the time they switch to a new procedure and do not return to E0. Because the final diagnosis is not necessarily the correct diagnosis, *Dif* is a useful a response variable for modeling operator progression through a scenario in dynamic HRA, but *Dif* should not be treated as a proxy for success or failure. As with *Trip*, *Dif* is a relative measure, normalized by dividing by the mean diagnosis time for the scenario.

In addition to the effects of simulator bias, *Trip* and *Dif* are affected by two factors. The first source of variation is the scenario itself—some scenarios require less time than others to complete. Scenario variations are accounted for by normalizing the time for each scenario—instead of reporting the absolute times, we divide the initial values of *Trip* and *Dif* by the mean response times for each scenario. This allows us to compare response times from two different accident scenarios. Thus, *Trip* ranges from 0.45 to 1.62, and *Dif* from 0.47 to 2.33.

The second source of variation is Crew Characteristics, which are expected to be the greatest source of variation in *Trip* and *Dif,* with simulator bias triggers small fluctuations on top of the variations between crews. To capture variations due to variations in crew, we report four crew characteristics: *Quiz*, *Evaluation*, *Team* and *BaseStress*. The Crew Characteristics are labeled X5-X8 in the models below,

indicating that like the Experimental Manipulations, they are independent variables expected to impact the Bias Measure and the Response Times. The Crew Characteristics are defined in Table 6.

Table 6: Variables in the OSU NPP Simulator Experiments - Crew Characteristics

Variable Name	Description
Crew Characteristics *(Not controlled for in the experimental design; collected for both SRO and RO)*	
[X5] Team	The operator's self-reported Team Dynamics. Survey question: "How would you rate the operating atmosphere of your 2-person team?" (scaled response; 0 = "We communicate poorly and do not trust each other;" 10 = We work well together. Communication is excellent."). In the SEM, *Team* is normalized from 0 to 1.
[X6] BaseStress	*BaseStress* measures the operators' tendency to report higher (or lower) stress than their colleagues. *BaseStress* is equal to the operator scenario normalized stress over all of the operator's simulator sessions (2 scenarios in 2014, 4 in 2015). The operator scenario normalized stress is the difference between the operator's reported stress in the scenario and the median reported stress in the scenario, minus the average difference between the reported stress and the median reported stress in the scenario, all divided by the standard deviation of the difference between the reported stress and the median stress for all the operators who participated in the scenario.
[X7] Evaluation	Researcher's 3-point rating of each operator's general ability to locate controls or follow the procedures in the simulator. 0 ~ below average, 0.5 ~ average, 1 ~ above average. Just over 50% of the students are rated 0.5.
[X8] Quiz	Grade from in-class quizzes. In the SEM, *Quiz* is a percent.

2.1.1 Data Collection

The data collection process used in this experiment loosely follows the six step process suggested by Spurgin and Petkov [15] for data collection in simulator studies: [Step 1] Scenarios were selected based on design-basis accidents and feasibility of operator response using Emergency Operating Procedures (EOPs). [Step 2] Researchers tested each scenario before using them in a study, and researchers and undergraduate research assistants tested data collection tools before using them in the experiments (see Table 7). [Step 3] Scenario runs were conducted with student operators. [Step 4] For debriefing, operators watched a video of their performance and reviewed the scenario, then completed the data collection instruments. [Step 5] Scenario context is determined in Step 1, and analysis of all the scenarios is completed following each iteration of the experiment. [Step 6] Data collection tools were updated after each iteration of the experiment.

Two preliminary studies have been conducted for this research (summarized in Table 7). In each study, students participated in a Nuclear Power Plant Systems and Operations course. As part of the course, students trained on the OSU NPP Simulator Facility with an emphasis on responding to design-basis accidents.

In the simulator experiments, students worked in two-person crews. Each crew had a Senior Reactor Operator (SRO) responsible for reading procedures and directing control room activities; and a Reactor Operator (RO) responsible for reading control board indications and performing actions as directed by the SRO. Simulator sessions were recorded on video. After each simulator session, operators completed a post-scenario survey to assess the bias during the session.

Table 7: OSU NPP Simulator Experiments

Year	Crews	Summary
2014	14 crews: 14	*Between subjects design*

[16], [17]	students, each in 2 crews with different partners	Each student completes a simple and complex version of the SGTR. Crews are randomly assigned; students do not have the same partner for the two scenarios. Scenarios: • SGTR (simple); larger SGTR + MSIV stuck open (complex)
2015	8 crews: 16 students, each in 1 crew with the same partner throughout	*Mixed subjects design* 2 phases: practice (2 scenarios, ungraded) and exam (2 scenarios, graded). 2 treatments: Simple (4 crews, only simple scenarios) and Complex (4 crews, simple and complex scenarios): • SGTR (simple); SGTR + SLB (complex) • LOCA (simple); LOCA preceded by stuck open Pressurizer PORV (complex) Treatment 0, Practice: SGTR, LOCA; Exam: LOCA, SGTR Treatment 1, Practice: SGTR, Complex LOCA; Exam: LOCA, Complex SGTR

Data from the 2014 and 2015 experiments are used in this analysis. Each scenario's data are averaged between the SRO and RO. Data collected in each iteration of the experiment include:

- Operator background survey responses
- Course grades and class work from the NPP Operations course for each student
- Post-scenario surveys following each simulator scenario
- Video/audio recordings of each scenario
- Dynamic PSF assessments of each scenario
- Plant parameters, alarm logs, and operator action logs recorded by the simulator software.

Data from 41 simulator scenarios are used in the analysis: 32 scenarios from the 2015 Experiment and nine from the 2014 Experiment.[3] For each scenario, both the Senior Reactor Operator (SRO) and the Reactor Operator (RO) contribute data. These data are averaged to estimate the overall scenario value. Data in include 8 scenario conditions (X1-X8; Table 2 and Table 6), five Bias Measures (M1-M5; Table 4) and two response times (R1-R2; Table 5). Bias Measures are ordinal data. Most are reported on a scale of 0-10. To consolidate the data into meaningful categories, the collected data are sorted into approximately five levels using the *Hist()* function in the statistical software package R. Response times are measured in seconds and normalized to minimize the differences in time required by the different scenarios and highlight the effect of operator actions on timing in the scenarios.

2.2 Structural Equation Models

We use SEM to develop analytical models of simulator bias. We model two aspects of simulator bias: the factors that impact bias and bias's effect on operator performance. Operator performance can be assessed by a wide variety of measures; for simplicity, in this preliminary analysis we limit our analysis to time of operator response.

The primary appeal of SEM is that this analytical method estimates latent variables, that is, variables that are not directly measurable. The Latent Bias Model (Figure 3) uses the Bias Measures to estimate a new variable, Bias. This variable is interpreted as the effect of the simulator environment on the operator response.

[3] In the 2014 Experiment, students completed the post-scenario questionnaire online after they left the simulator facility. Students submitted questionnaires for only nine of the observed scenarios. To improve the low response rate, we asked 2015 students to complete the questionnaire before leaving the simulator facility.

However, our analysis found that several Bias Measures were not significant factors in the latent Bias variable, but that these Bias Measures had a significant impact on the operator response. This suggests that multiple Bias effects are active in the scenario. Unfortunately, there are not enough Bias Measures to estimate multiple latent variables. Instead, we introduce a second model, the Bias SEM model (Figure 4), which illustrates a second application of SEM: estimating the relationships between measured variables without estimating latent variables. The Bias SEM estimates the relationships between the scenario conditions, the Bias Measures, and the operator response.

Together, the Latent Bias Model and the Bias SEM illustrate two related applications of SEM. The Latent Bias Model provides a quantitative estimate of simulator bias, but some information is lost because all the Bias Measure are assumed to contribute to a single effect. The Bias SEM Model retains more information, but is less convenient because it does not produce a single, defined measure of simulator bias.

Section 2.3.1 provides a brief overview of SEM for readers who are new to SEM while introducing two preliminary Bias SEMs. Section 2.3.2 describes the process used to generate the final simulator bias models discussed in Section 3. Table 8 lists commonly used SEM terminology.

Table 8: SEM terminology.

SEM Term	*Definition*
Independent Variable	A variable that is not a function of other variables in the model. Also called endogenous variables. Scenario Conditions and Crew Characteristics in Figure 3 are independent.
Dependent variable	A variable that is modeled as a function of other variables in the model. Also called exogenous variables. Dependent variables may contribute to other variables in the model; Bias Measures and Response Times in the Path Analysis Model in Figure 1 are dependent variables.
Latent variable	An estimated (unmeasured) variable, associated with at least two indicators. In Figure 3, *Bias* is the latent variable estimated in the Latent Variable Model.
Indicator	A measured variable used to estimate a latent variable. Indicators are dependent variables. Bias Measures in the Latent Bias Model are indicators of the latent variable Bias.
Factor loadings	The coefficient estimating the relationship between the latent variable and an indicator variable. For each latent variable, one indicator is anchored to the latent variable—i.e., the factor loading for that indicator is fixed to 1.
Path Analysis Model	A model with no latent variables
Measurement Model	Model of a latent variable, associated indicators, and factor loadings.

2.2.1 Introduction to SEM

Figure 2 illustrates the process used to estimate the unknown parameters in the structural model from two primary inputs: a set of **observed variables** and a proposed **structural model** of the relationships between variables [18]. A secondary input is the **observed covariance matrix**, that is, a matrix of the covariances between the observed variables. An algorithm suggests values for the unknown model parameters in the structural model (**Suggested Parameters**). An **estimated covariance matrix** is calculated from the suggested model parameters. A **likelihood estimator** is used to compare the estimated covariance matrix to the **observed covariance matrix**. Based on these results, the algorithm suggests a new set of model parameters, attempting to increase the likelihood of the estimated covariance matrix. This process is repeated, with the algorithm selecting progressively better suggested estimates until the model converges [19].

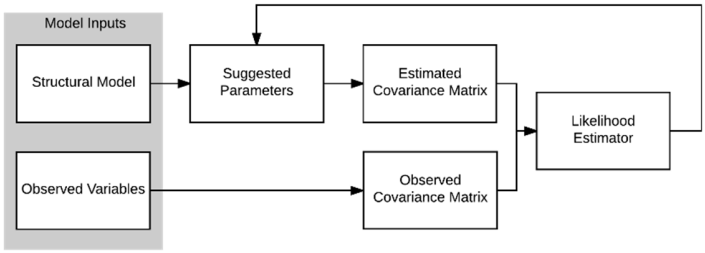


Figure 2: The iterative SEM analysis process.

Let us use Figure 2 to develop the Latent Bias Model. The **structural model** specifies the hypothesized relationships between variables. One of the strengths of SEM is the ability to assess latent variables—that is, variables that are not directly measured. Structural models have two parts: a path analysis model of the relationships between variables, and a measurement model that relates the latent (unobserved) variable to its associated (measured) indicators. Structural models are often represented graphically; Figure 3 organizes the variables from Table 2, Table 4, Table 5, and Table 6 into a preliminary structural model of *Bias.*

The preliminary Latent Bias Model treats *Bias (B)* as a function of the Conditions (*X1-X8*), and as a predictor of the response time variables (*R1* and *R2*). The measurement model in Figure 3 shows *Bias* measured by the five Bias Measures, *M1-M5*: *Unreal, Prepared, Likely.For, Likely.Against* and *Worried.*

Following SEM convention, Figure 3 uses ovals to indicate unobserved variables and rectangles to indicate measured variables. In addition to the unobserved variable *Bias*, the model also estimates variance or measurement errors for the Bias Measures (*Em1-Em4*) as well as the response variables (*Er1, Er2*).

Arrows show the presumed direction of effects; variables that are at the end of an arrow are dependent variables, while variables that are only the source of arrows are independent variables.

To translate the graphical representation into linear equations, we define three vectors of observed variables:[4]

- Response vector $\boldsymbol{R} = [Trip, Dif]$
- Conditions vector $\boldsymbol{X} = [Trt1, Exam, Fami, Yr2015, Team, BaseStress, Evaluation, Grades]$
- Measurement vector $\boldsymbol{M} = [UnReal, Prepared, Likely, Worried]$

Using *B* to represent the latent variable *Bias*, we define four new variables to describe the relationships between ***R, X*** and ***M:***

- $\boldsymbol{\Gamma}$ [Gamma], a 2 x 8 matrix of regression coefficients of the conditions ***X*** on the response times ***R***
- $\mathbf{A}$ [Alpha], an eight-element vector of regression coefficients of ***X*** on *B*
- $\boldsymbol{\Lambda}$ [Lambda], a four-element vector of factor loadings between *B* and ***M***
- $\mathbf{B}$ [Beta], a two-element vector of regression coefficients of the *B* on ***R***

[4] This nomenclature is loosely based on **Invalid source specified.**. Vectors and matrices are written in bold to distinguish them from individual coefficients or variables.

We also define two variance vectors:

- $\boldsymbol{E_m}$, the four-element measurement error vector in the measurement model, ***M***
- $\boldsymbol{E_r}$, the two-element measurement error vector tied to the response times, ***R***

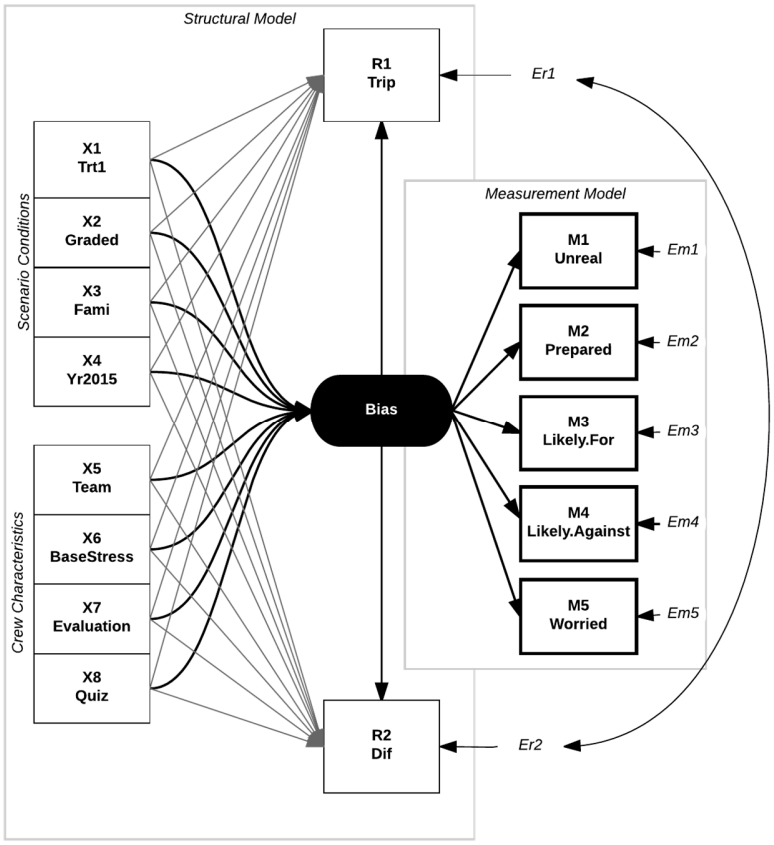


Figure 3:The preliminary OSU Latent Bias Model.

Using these defined variables, we can write the SEM in Figure 3 as a set of equations. Equations 1a and 1b are the structural model, and Equation 1c is the measurement model:

$$B = \boldsymbol{A} \cdot \boldsymbol{X} \tag{1a}$$

$$\boldsymbol{R} = \boldsymbol{\Gamma X} + \mathbf{B}B + \boldsymbol{E_r} \tag{1b}$$

$$\boldsymbol{M} = \boldsymbol{\Lambda}B + \boldsymbol{E_m} \tag{1c}$$

The **Estimated Parameters** [19] include

- Regression coefficients in the path analysis model ($\mathbf{A}, \boldsymbol{\Gamma}$)
- Factor loadings between the latent variable bias and the Bias Measures in the measurement model ($\boldsymbol{\Lambda}$)

  - The researcher must define at least one factor loading per latent variable to provide a base for scale; the other factor loadings are estimated relative to the fixed factor loading. In this model, the factor loading for *Unreal* is defined as $\Lambda_1 = 1.0$, and Λ_2, Λ_3 and Λ_4 are allowed to vary. We selected *Unreal* as the anchor for *Bias* because *Unreal* is an explicit measure of perceived operator bias in the simulator.
- Variance or measurement errors of observed variables ($\boldsymbol{E_m}, \boldsymbol{E_r}$)
- Covariances between independent variables, i.e. between ***X***, $\boldsymbol{E_m}$ and $\boldsymbol{E_r}$ (covariances between dependent variables are calculated from $\mathbf{B}$ and $\mathbf{\Lambda}$ values).

To select the best model, the **estimated covariance matrix** is compared to the **observed covariance matrix** from the observed variables. A **likelihood estimator** is used to select between possible models as the algorithm proposes new sets of estimated parameters. We use the robust Weighted Least Squares with the mean- and variance-adjusted test statistic (WLSMV) in the R package developed by The lavaan Project [20] to estimate model parameters. The WLSMV estimator is lavaan's default estimator for models that include categorical dependent variables such as the Bias Measures in our model.

2.2.1.1 Bias Path Analysis Model

Similar to the preliminary Latent Bias Model, the preliminary Bias Path Analysis Model is shown in Figure 4. We include the Bias Path Analysis Model to attempt to capture the different bias effects in the Bias Causal Model: *Likely.For, Likely.Against* and *Prepared* are measures of the Prominent Hypothesis Bias, *Worried* measures the Incentive Effect, and *Unreal* is a measure of overall bias. Ideally, each of these biases would be included as a separate latent variable in the model, but there are too few Bias Measures to estimate separate bias effects (each latent variable requires at least two indicators, preferably three or more). In lieu of a full latent variable model estimating multiple latent biases, we use the Bias Path Analysis Model in Figure 4 to examine multiple biases that are active in the OSU NPP Simulator Experiment.

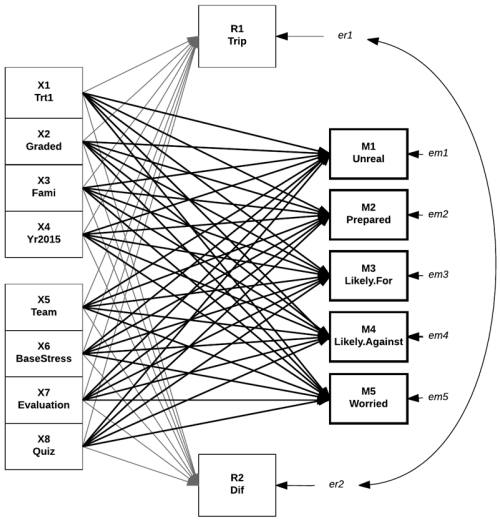


Figure 1: The Preliminary Bias Path Analysis Model

As with the Latent Bias Model, the Bias Path Analysis Model model can be written as a set of linear equations. The variables ***X***, ***M*** and ***R*** remain the same; the vectors in the Latent Bias Model become matrices in the Bias Path Analysis Model. The Bias Path Analysis Model coefficients are named with lower case letters to differentiate them from the Latent Bias Model coefficients:

- $\boldsymbol{\gamma}$ [gamma], a 2 by 8 matrix of regression coefficients between ***X*** and ***R***
- $\boldsymbol{\alpha}$ [alpha], a 4 by 8 matrix of regression coefficients between ***X*** and ***M***
- $\boldsymbol{\beta}$ [beta], a 2 by 4 matrix of regression coefficients between ***M*** and ***R***
- $\boldsymbol{e_m}$, a four-element vector of estimated variances for ***M***

- $\boldsymbol{e_r}$, a two-element vector of estimated variances for $\boldsymbol{R}$

The structural model equations for the Bias Path Analysis Model are therefore

$$\boldsymbol{M} = \boldsymbol{\alpha X} + \boldsymbol{e_m} \tag{2a}$$

$$\boldsymbol{R} = \boldsymbol{\gamma X} + \boldsymbol{\beta M} + \boldsymbol{e_r} \tag{2b}$$

Note that the Bias Path Analysis Model does not include a measurement model, because no latent variables are present.

2.2.1.2 SEM Mediation

In both the Latent Bias Model and the Bias Path Analysis Model, Scenario Conditions and Crew Conditions have direct and indirect effects on *Trip* and *Dif.* The indirect effect is the effect of these conditions through *Bias* (Figure 3) or the Bias Measures (Figure 4). This is known as mediation [21]: simulator bias mediates the effect of experimental conditions. Figure 5 diagrams the direct and indirect (i.e. mediated) effects.

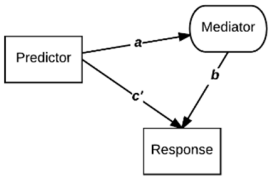


Figure 2: Mediation in SEM

In Figure 5, *c'* is the direct effect of the predictor on the response variable, and *ab* (that is, a multiplied by b) is the indirect effect. The total effect, *c,* is the sum of the direct and indirect effect,

$$c = c' + ab \tag{3}$$

For example, consider the effect of *Team* on *Trip* in Figure 3. *Team* is the predictor variable, *Bias* is the mediator, and *Trip* is the response variable. Using the nomenclature from Equation 1, the variables in Figure 5 are

- $a = \mathrm{A}_5^s$, the standardized regression coefficient[5] for *Team* (*X5*) on *Bias (B)*
- $b = \mathrm{B}_1^s$, the standardized regression coefficient for *Bias (B)* on *Trip* (*R1*)
- $c' = \Gamma_{51}^s$, the standardized regression coefficient for *Team* (*X5*) on *Trip* (*R1*)

To calculate the total effect of *Team* on *Trip, c,* we define ΔX_5 as a change in *Team* and ΔR_1 as the resulting change *Trip* because of the change in *Team.* We then calculate $\Delta R_1 = (ab + c')\Delta X_5$, i.e.

$$\Delta R_1 = (A_5^s B_1^s + \Gamma_{51}^s)\Delta X_5 \tag{4}$$

One aspect of the analysis is discerning how much (if at all) a particular effect is mediated by simulator bias. In this study, the amount of mediation corresponds to how much simulator bias affects the outcome. Variables can be fully mediated, partially mediated, unmediated, or inconsistently mediated:

- **Effect:** the predictor variable has an effect on the response variable
  - **Full Mediation:** The total effect is accounted for indirectly by simulator bias ($ab = c, c' = 0$)

[5] Standardized coefficients are calculated in units of standard deviation. For example, a standard deviation change in *Team* (*X5*) will change *B* by α_5^s standard deviations of *B*. Standardized coefficients are useful for comparing the impact of variables that have different. In this analysis, standardized coefficients are represented by a superscript *s* ($\boldsymbol{\alpha^s}, \boldsymbol{\beta^s}, \boldsymbol{\gamma^s}$, etc.).

  - **Partial Mediation:** The factor has both a direct and an indirect effect ($ab + c' = c, ab \neq 0, c' \neq 0$).
    - **Inconsistent mediation:** Variables with indirect effects that counteract the direct effect (ab, c' have opposite signs; $ab \times c' < 0$)—the indirect effect reduces the overall effect the variable would otherwise have on the response variable [22].
- **No Mediation:** The factor has no indirect effect and is fully accounted for without considering simulator bias ($ab = 0, c' = c$)
- **No Effect:** This effect is insignificant and not included in the final model ($ab \sim 0, c' \sim 0$; note that a or b may be significant, but together they are not important in the model).

Table 9: Notation frequently used in this paper.

Notation	Description
SEM NOTATION	
R1, R2	Response Times *Trip* and *Dif*
M1-M4	Bias Measures
X1-X8	Scenario Conditions
$\mathrm{A}_i, \alpha_{ij}$	SEM regression coefficients between X_i and the latent variable *Bias (B)* or M_j (upper case: Latent Bias Model; lower case: Bias Path Analysis Model)
$\alpha^s_{ij}, \beta^s_{ij}, \ldots$	Standardized coefficients
Λ_i	Lambda: SEM factor loadings between the latent variable *Bias* (B) and the *Bias Measure* M_i
B_i, β_{ij}	Beta, beta: SEM regression coefficients between Bias (B) or Bias Measure M_j on response time R_i
Γ_{ij}, γ_{ij}	Gamma, gamma: SEM regression coefficient between scenario condition X_j and response time R_i
E, e	SEM estimated variance of observed variables
a	a: In a mediated model, the effect of the predictor variable on the mediating variable (α^s)
b	b: In a mediated model, the effect of the mediating variable on the response variable (β^s)
ab	ab: $a \times b$, In a mediated model, the indirect effect of the predictor on the response variable
c	c: In a mediated model, the total effect of the predictor on the response variable, *ab+c*
c'	c-prime: In a mediated model, the direct effect of the predictor on the response variable (γ^s)
OTHER VARIABLES USED IN THIS ANALYSIS	
p	The p-value. For the model parameter estimates, low p-values ($p < 0.2$) indicate we expect the parameter is significant in the model. (Note: SEM convention uses p to denote the number of parameters in the model)
ε	epsilon: The Root Mean Square Error of Approximation (RMSEA) for the SEM
d	The degrees of freedom in the model
n	The sample size
λ	lambda: The noncentrality parameter in the noncentral χ^2 distribution
H_0, H_a	The null and alternative hypotheses. In this analysis, the null hypothesis is that the model is a good fit to the data, i.e., $\varepsilon < 0.05$; the alternative hypothesis is that the model does not fit the data, i.e., $\varepsilon > 0.08$.
α	alpha: The maximum threshold for the probability that the model is not a good fit for the data.
π	pi: The power of the test, i.e., the probability of not making a Type II Error.
r	In SEM, the ratio of the number of indicators to the number of estimated parameters in the model.

2.2.2 Final Model Selection

All possible direct and indirect effects are included in the preliminary models (Figure 3 and Figure 4). This is because we can provide reasonable explanations for how all of the independent variables might impact each of the dependent variables.

We expect that the effects of Scenario Conditions (*X5-X8*) on Response Times (*R1,R2*) will be largely mediated by *Bias*; that is, that many of the direct effects of Scenario Conditions on Response Times will be negligible.[6] However, we retain the direct effects of these variables in the model because we do not have a strong theoretical reason to exclude direct effects of Scenario Conditions on Response Times from the preliminary models. The final models are developed by removing insignificant links in the model; this process is illustrated in Figure 6. Model coefficients for the preliminary model are generated using the WLSMV estimator in lavaan. Then, variables with low standardized coefficients are removed from the model. These variables have a negligible effect on the response variable. In this analysis, we retain all variables with a standardized coefficient greater than 0.1. Negligible links between variables are removed iteratively; links with the smallest standardized coefficients are removed (i.e. coefficients are set equal to zero in the model), then the model is re-evaluated and again the variables with the smallest standardized coefficients are removed from the model. This process is repeated until all of the estimated standardized coefficients in the model are 0.1 or greater.

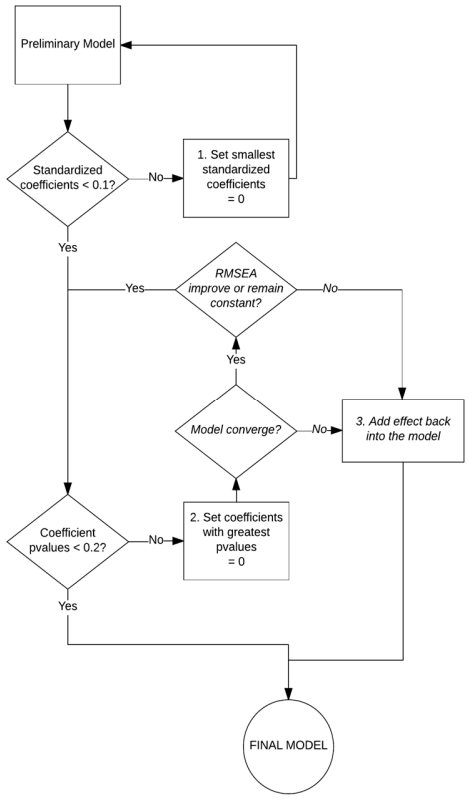


Figure 3: SEM Model Development Process

Finally, coefficients with high p-values (greater than 0.2) are removed from the models. In SEM, p-value can be interpreted similarly to p-values for coefficients in linear regression analysis. Low coefficient p-values suggest that the effect is significant; high p-values indicate that the effect is likely to be insignificant.

This continues until all of the coefficient p-values are less than 0.25 unless the model fails to converge after a coefficient is removed, or unless the model Root Mean Square Error of Approximation (RMSEA, a measure of model fit—see Section 3.3) increases when the coefficient is removed. In either case, the coefficient is returned to the model (**Step 3**), and the model revision process is complete. The results of this analysis are shown in Section 3.

3 RESULTS

The final versions of the Latent Bias Model and the Bias Path Analysis Model are introduced in Section 3.1, with diagrams of the interactions between model elements and a brief discussion of bias

[6] Recall that *Trip* and *Dif* are scenario-normalized response times, so variations between scenarios are minimized in the analysis.

interactions and their impact on *Trip* and *Dif* for each model. Section 3.2 summarizes the mediated effects in the models.

Section 3.3 introduces two approaches to model validation that can be used to assess models that are developed from experienced operator data, and Section 3.4 estimates the minimum sample size needed to assess these models with a high degree of confidence.

3.1 BIAS STRUCTURAL EQUATION MODELS

Two Bias SEMs are developed: the Latent Bias Model and the Bias Path Analysis Model. As stated above, the coefficients in both models are estimated using data collected from student operators; the numeric results are provided to illustrate how the models can be developed but are not meant to represent experienced operators.

3.1.1 Latent Bias Model

The Latent Bias Model estimates the latent variable *Bias*. Figure 7 is the Latent Bias Model diagram. The estimated coefficients are reported in Table 10, along with the standardized coefficient and the coefficient p-value. The p-value is the probability that the effect is negligible (i.e. equal to zero).

In the Latent Bias Model, the Bias Measures *Likely.For* and *Likely.Against* are removed from the model because they are not significant measures of the latent variable *Bias*.

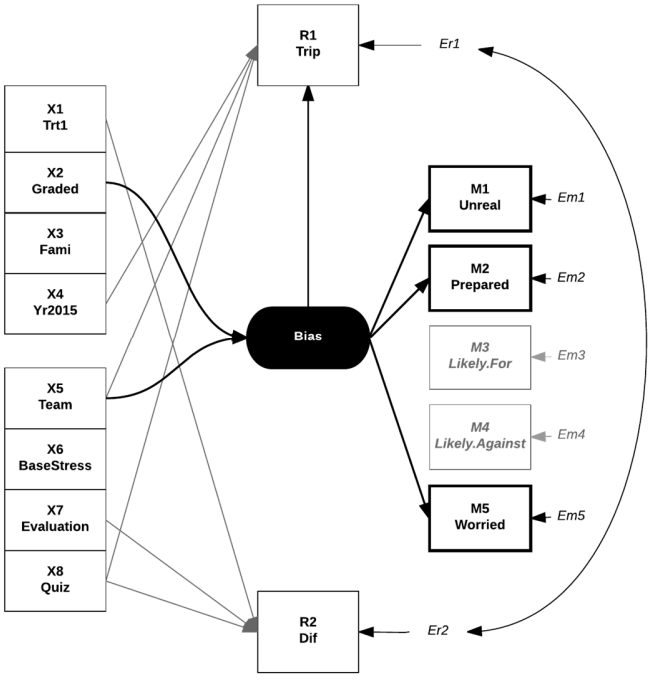


Figure 7: The Latent Bias Model estimates the latent variable, Bias. *The Bias Measure* Likely.Dist *and the crew characteristic* BaseStress *are not included in the model because they are found to be insignificant in the process described in Section 2.3.2.*

Table 10: The Latent Bias Model estimated coefficients. Standardized coefficients represent the change in response variable per change in predictor variable in units of standard deviation. P-values are the probability the effect is negligible; bootstrap p-values are estimated through bootstrap resampling.

Latent Bias Model		Coefficient			Standardized Coefficient			Pvalue
Bias =~ ΛM	Unreal	Λ_1	=	1.000	Λ_1^s	=	1.014	–
	Prepared	Λ_2	=	−0.378	Λ_2^s	=	−0.494	0.086
	Worried	Λ_5	=	0.173	Λ_5^s	=	0.235	0.164
Bias ~ AX	Graded	A_2	=	−1.296	A_2^s	=	−0.464	0.016
	Yr2015	A_4	=	−0.853	A_4^s	=	−0.259	0.227
	Team	A_5	=	−5.216	A_5^s	=	−0.535	0.023
Trip~ $B_1 + \Gamma_1 X$	Bias	B_1	=	0.05	B_1^s	=	0.214	0.141
	Yr2015	Γ_{14}	=	−0.467	Γ_{14}^s	=	−0.612	0.003
	Team	Γ_{15}	=	−0.852	Γ_{15}^s	=	−0.377	0.037
	Quiz	Γ_{18}	=	−2.417	Γ_{18}^s	=	−0.596	0.022
Dif ~ $B_2 + \Gamma_2 X$	Trt1	Γ_{21}	=	0.258	Γ_{21}^s	=	0.363	0.069
	Stress	Γ_{26}	=	0.108	Γ_{26}^s	=	0.227	0.244
	Evaluation	Γ_{27}	=	−0.668	Γ_{27}^s	=	−0.555	0.07

In the Latent Bias Model, *Bias* is measured by three of the four Bias Measures: *Unreal, Prepared,* and *Worried.*

3.1.1.1 Effects of Manipulations on Simulator Bias in the Latent Bias Model

Bias is a function of *Graded, Yr2015* and *Team*, indicating that simulator bias decreases in the exam session and in 2015, and *Bias* increases if team dynamics are poor. This matches expectations: *Graded* is expected to decrease *Bias*, because the exam session is meant to mimic the pressure and urgency of operating a nuclear power plant. *Yr2015* is essentially a measure of training; in 2015, students had more time in the simulator and more experience responding to accident scenarios. This indicates that improved training reduced the artificiality of the simulator experience. Finally, poor team dynamics indicated by low *Team* makes crews more aware of the artificial environment.

Bias is not a function of the other manipulations (*Trt1, Fami*), which might suggest that these manipulations do not impact simulator bias. However, remember that, due to sample size constraints, only one latent variable is included in the Latent Bias Model, and that variable (*Bias*) is anchored to *Unreal.* The variables *Trt1* and *Fami* may be influencing bias effects that are not related to *Unreal* and are therefore not captured in the model. In other words, *Bias* captures only a part of the bias effects. This interpretation is supported by the Bias Path Analysis Model.

3.1.1.2 Effects of Bias on Operator Response in the Latent Bias Model

Similarly, *Trip* is dependent on *Bias*, but *Dif* is not. Again, this does not mean that *Dif* is not susceptible to simulator biases. Instead, we believe *Bias* does not capture all of the effects of simulator bias. *Dif*'s independence from *Bias* in the Latent Bias Model suggests that treating *Bias* as a single variable is too simplistic. For example, the bias measured by *Likely.Against* may have a significant impact on *Dif*, but because this effect is not the same as the effect measured by *Unreal*, the *Likely.Against* effect is dropped from the Latent Bias Model and *Likely.Against*'s impact on *Dif* is not captured. The Bias Path

Analysis Model (Section 3.1.2) supports this interpretation of the results. In the Bias Path Analysis Model, *Dif* is a function of *Likely.Against* and *Worried*, while *Trip* is a function of *Unreal* and *Likely.Against.*

3.1.2 Bias Path Analysis Model

The Bias Path Analysis Model treats the Bias Measures as proxies for their associated biases. Figure 8 is the Bias Path Analysis Model diagram. The estimated coefficients for the Bias Path Analysis Model are reported in Table 11, along with the standardized coefficient, the coefficient p-value, and a bootstrap p-value.

In the Bias Path Analysis Model, two Bias Measures are removed from the model because they do not have a significant effect on *Trip* or *Dif*: *Prepared* and *Likely.For*.

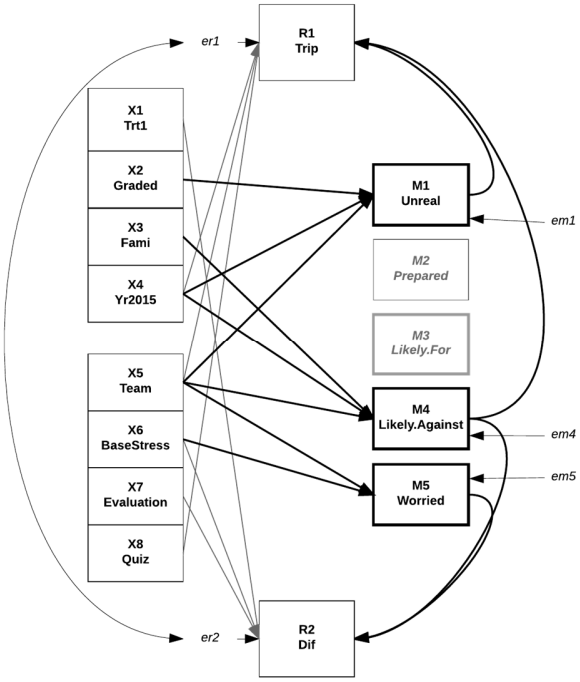


Figure 8: The Bias Path Analysis Model. Bias Measures Prepared *and* Likely.For *are dropped from the model.*

Table 11: The Bias Path Analysis Model estimated coefficients. Standardized coefficients represent the change in response variable per change in predictor variable in units of standard deviation. P-values are the probability the effect is negligible; bootstrap p-values are estimated through bootstrap resampling.

Bias Path Analysis Model		**Coefficient**	**Standardized Coefficient**	**pvalue**
Unreal ~ $\alpha_1 X$	*Graded*	$\alpha_{12} = -1.12$	$\alpha_{12}^s = -0.434$	0.046
	Yr2015	$\alpha_{14} = -0.933$	$\alpha_{14}^s = -0.307$	0.238
	Team	$\alpha_{15} = -4.805$	$\alpha_{15}^s = -0.533$	0.03
Likely.Against ~ $\alpha_4 X$	*Fami*	$\alpha_{43} = -0.119$	$\alpha_{43}^s = -0.274$	0.233
	Yr2015	$\alpha_{44} = -0.172$	$\alpha_{44}^s = -0.498$	0.042
	Team	$\alpha_{45} = -0.513$	$\alpha_{45}^s = -0.502$	0.07

Bias Path Analysis Model		Coefficient	Standardized Coefficient	pvalue
Worried ~ $\alpha_5 X$	*Team*	$\alpha_{55} = -2.631$	$\alpha^s_{55} = -0.324$	0.225
	BaseStress	$\alpha_{56} = 0.532$	$\alpha^s_{56} = 0.343$	0.176
Trip ~ $\beta_1 R + \gamma_1 X$	*Unreal*	$\beta_{11} = 0.071$	$\beta^s_{11} = 0.279$	0.056
	Likely.Against	$\beta_{14} = -0.493$	$\beta^s_{14} = -0.219$	0.116
	Yr2015	$\gamma_{14} = -0.546$	$\gamma^s_{14} = -0.705$	0.001
	Team	$\gamma_{15} = -1.035$	$\gamma^s_{15} = -0.451$	0.015
	Quiz	$\gamma_{18} = -2.415$	$\gamma^s_{18} = -0.587$	0.024
Dif ~ $\beta_2 R + \gamma_2 X$	*Likely.Against*	$\beta_{24} = 0.421$	$\beta^s_{24} = 0.172$	0.103
	Worried	$\beta_{25} = -0.067$	$\beta^s_{25} = -0.217$	0.195
	Trt1	$\gamma_{21} = 0.223$	$\gamma^s_{21} = 0.313$	0.158
	BaseStress	$\gamma_{26} = 0.147$	$\gamma^s_{26} = 0.309$	0.107
	Evaluation	$\gamma_{27} = -0.682$	$\gamma^s_{27} = -0.564$	0.064

In the Bias Path Analysis Model, *Unreal* has the same predictors that *Bias* has in the Latent Bias Model: *Graded, Yr2015* and *Team*. This highlights the influence of *Unreal* on the estimate of *Bias* in the Latent Bias Model.

3.1.2.1 Effects of the Manipulations on Simulator Bias in the Bias Path Analysis Model

Trt1: *Trt1* is not a predictor for any of the Bias Measures. We expected *Trt1* to manipulate prominent hypothesis bias (measured by *Prepared, Likely.For* and *Likely.Against*). As *Trt1* is not a significant predictor, it is likely that the effects *Fami* overpower the impact of *Trt1* in the model—remember that only *Trt1=1* crews encountered scenarios with low *Fami*.

Graded: Students in *Graded=1* scenario report a decrease in *Unreal,* which measures *Cavalier Behavior*. *Graded* is included in the experiment to trigger the *Incentive Effect*, which is expected to reduce *Cavalier Behavior*. The effect of *Graded* on *Unreal* supports this interpretation of the interacting biases *Incentive Effect* and *Cavalier Behavior.*

Contrary to expectations, *Graded* is not a predictor for *Worried*, which is meant to measure the *Incentive Effect*. Instead, *Worried* is a function of *Team* and *BaseStress*. This indicates that *Worried* is not a good measure of *Incentive Effect,* which can be seen indirectly through *Graded*'s effect on *Unreal*. The strength of the incentive in graded sessions (*Graded = 1*) is not strong enough to be seen over the impact of crew characteristics on operator's anxiety about their grades.

Fami: *Fami* is meant to manipulate the *Prominent Hypothesis Bias*, which is measured by *Prepared, Likely.For* and *Likely.Against*. We cannot see the impact of *Fami* on *Prepared* or *Likely.For* because these measures are not included in the model (they are not significant predictors for on *Trip* or *Dif*). As expected, *Likely.Against* decreases as *Fami* increases: the more familiar the scenario, the less likely students are to expect other accidents or events.

Yr2015: *Yr2015* is a measure of training, an incidental bias manipulation introduced by the evolution of the experiment over the two years of data collection. We had no theoretical expectations of how *Yr2015*

would impact simulator bias. In the Bias Path Analysis Model, *Yr2015* is a predictor for *Unreal* and *Likely.Against*. The tie between *Yr2015* and *Unreal*—like the tie between *Yr2015* and *Bias* in the Latent Bias Model—highlights the increase in perceived reality of the simulator environment with additional training and knowledge of the accidents. The decrease in *Likely.Against* in 2015 reflect the fact that some students in 2014 had strong, incorrect expectations about the coming scenario, while students in 2015 were more likely to have a strong correct hypothesis or to report no expectations about the scenario.

3.1.2.2 Effects of Bias on Operator Response in the Bias Path Analysis Model

In the Bias Path Analysis Model, *Trip* is a function of the Bias Measures *Unreal* and *Likely.Against*, while *Dif* is a function of *Likely.Against* and *Worried*. Although *Worried* is meant to evaluate *Incentive Effects*, from the discussion above we see that *Worried* is a function of crew characteristics rather than the phase of the experiment.

Unreal: From the Bias causal model, *Unreal* represents *Cavalier Behavior*. The Bias Path Analysis Model therefore suggests that *Cavalier Behavior* (through *Unreal*) increases the time for operators to trip the reactor (*Trip*) but has no effect on the time to diagnose the accident after the reactor trips (*Dif*).

Likely.Against: *Likely.Against* measures an operator's *Prominent Hypothesis*, specifically, the strength of false expectations about the scenario. As might be expected, *Dif* increases with *Likely.Against*—the stronger an operator's false expectations, the longer it takes to diagnose the accident and leave E0 for the correct accident-specific procedure. Somewhat counterintuitively, *Trip* decreases as *Likely.Against* increases. This can be explained by the fact that operators trip the reactor after any severe accident, often before they have a strong understanding of the cause of the accident. In other words, expectations of any accident would decrease the time to trip the reactor. *Dif*, on the other hand, measures the time to identify the correct response to the specific accident, and will naturally increase if the operators have a mistaken initial hypothesis.

Worried: In the Bias Path Analysis Model, *Dif* decreases as *Worried* increases. One possible explanation for this effect is that students who are worried about their performance rush through the procedures in an effort to look competent and capable, while students who are not worried about the grades feel comfortable taking more time to respond to the accident. Also, *Worried* decreases as *Team* improves—in some instances, crews with better team dynamics take more time to discuss decisions before moving through the procedures, while in crews with poor team dynamics all decision making is made by the SRO without consulting with the RO.

3.2 MEDIATED EFFECTS

Table 12 summarizes the indirect effects in the Latent Bias Model. The Indirect Effect (i.e. *ab* in Figure 5) is the effect of the scenario condition on the response variable that is mediated by *Bias*. The Indirect Effect is the standardized coefficient of the effect of the condition on *Bias* multiplied by the standardized coefficient for *Bias*'s effect on the response variable.

Partial Mediation in Table 12 is the indirect effect divided by the total effect. This variable shows the relative strength of the mediation.

There are no indirect effects acting on *Dif* because *Bias* is not a predictor of *Dif* in the Latent Bias Model.

For *Trip*, *Graded* is fully mediated by *Bias*. Without the simulator environment, the model indicates that *Graded* would have no effect on *Trip*. *Team* is partially mediated by *Bias*; simulator bias accounts for approximately twenty-five percent of *Team*'s effect on *Trip.* Outside the simulator environment, we would expect *Team* to be less influential on *Trip*.

Table 12: Mediation in the Latent Bias Model

Trip ~	**Mediation**	**Standardized Coefficient *(ab)***		**pvalue**	**Partial Mediation (ab/c)**		**Pvalue**
Graded	Complete	$A_2^s B_1^s$	−0.099	0.133			
Yr2015	Partial	$A_4^s B_1^s$	−0.056	0.342	$\frac{A_4^s B_1^s}{A_4^s B_1^s + \Gamma_{14}^s}$	*0.083*	*0.346*
Team	Partial	$A_5^s B_1^s$	−0.115	0.205	$\frac{A_5^s B_1^s}{A_5^s B_1^s + \Gamma_{15}^s}$	*0.233*	*0.238*

In the Bias Path Analysis Model, many of the mediated effects (Table 13) are inconsistent—that is, the simulator bias counteracts the direct effect of the condition on the response variable.

For example, Table 11 shows that *BaseStress's* direct effect is to increase *Dif* ($\gamma_{26}^s > 0$). However, the mediated effect of *BaseStress* through *Worried* is to decrease *Dif* (from Table 13, $\alpha_{56}^s \beta_{25}^s < 0$). Therefore, this model suggests that an operator's tendency towards stress and anxiety will have a stronger effect in the control room than in the simulator environment.

Table 13: Mediation in the Bias Path Analysis Model.

Trip ~	**Mediation**	**Standardized Coefficient *(ab)***		**pvalue**	**Partial Mediation (ab/c)**		**Pvalue**
Graded	Unreal, Complete	$\alpha_{12}^s \beta_{11}^s =$	−0.121	0.081			
Fami	Likely.Against, Complete	$\alpha_{43}^s \beta_{14}^s =$	0.06	0.348			
Yr2015	Unreal, Partial	$\alpha_{14}^s \beta_{11}^s =$	−0.086	0.24	$\frac{\alpha_{14}^s \beta_{11}^s + \alpha_{44}^s \beta_{14}^s}{\alpha_{14}^s \beta_{11}^s + \alpha_{44}^s \beta_{14}^s + \gamma_{14}^s} =$	−0.035	0.831
	Likely.Against, Inconsistent	$\alpha_{44}^s \beta_{14}^s =$	0.109	0.233			
Team	Unreal, Partial	$\alpha_{15}^s \beta_{11}^s =$	−0.149	0.121	$\frac{\alpha_{15}^s \beta_{11}^s + \alpha_{45}^s \beta_{14}^s}{\alpha_{15}^s \beta_{11}^s + \alpha_{45}^s \beta_{14}^s + \gamma_{15}^s} =$	0.079	0.764
	Likely.Against, Inconsistent	$\alpha_{45}^s \beta_{14}^s =$	0.11	0.274			
Dif ~	**Mediation**	**Standardized Coefficient *(ab)***		**pvalue**	**Partial Mediation (ab/c)**		**Pvalue**
Fami	Likely.Against, Complete	$\alpha_{43}^s \beta_{24}^s =$	−0.047	0.378			
Yr2015	Likely.Against, Complete	$\alpha_{44}^s \beta_{24}^s =$	−0.086	0.246			
Team	Likely.Against, Complete	$\alpha_{45}^s \beta_{24}^s =$	−0.087	0.255			
	Worried, Complete	$\alpha_{55}^s \beta_{25}^s =$	0.07	0.39			
Base Stress	Worried, Inconsistent	$\alpha_{56}^s \beta_{25}^s =$	−0.075	0.393	$\frac{\alpha_{56}^s \beta_{25}^s}{\alpha_{56}^s \beta_{25}^s + \gamma_{26}^s} =$	−0.318	0.545

3.3 Model Fit & Sample Size Recommendations

To assess the validity of models such as the Latent Bias Model and the Bias Path Analysis Model, two aspects of the model should be examined: the overall fit of the model to the data, and the strength of the interactions between model elements (i.e. the estimated coefficients)

In Section 3.3.1, we examine the significance of each estimated effect in the model using bootstrap resampling. Overall, the bootstrap supports the initial model values.

In Section 3.3.2, we assess the overall model fit using the Root Mean Square Error of Approximation (RMSEA). We follow the consensus that an RMSEA of less than 0.05 represents a reasonable model fit [23]. Based on this threshold, in Section 3.3.3 we estimate the sample size needed to eliminate the possibility of an indeterminate RMSEA with a power of 80%.

Finally, in Section 3.4 we discuss how the recommended sample size changes if we wish to assess the strength of specific mediated effects in the analysis.

3.3.1 Bootstrap Validation of the Significance of Interactions Included in the Model

Bootstrap analysis is a common resampling method used to validate models. We use the bootstrap sampling approach outlined in [24] to determine the p-value for the model coefficients. First, 1,000 bootstrap samples of 41 data points each are drawn from the original sample. We fit the model to each sampled dataset and store the estimated coefficients for each model that converges. The model converges successfully for approximately half of the sample data sets, resulting in a set of approximately 500 estimates of each coefficient.

The pvalue for the coefficient is the probability that the effect is negligible. If the coefficient is positive, the bootstrap coefficient pvalue is simply the proportion of the bootstrap estimated coefficients that are less than zero. Similarly, if the coefficient is negative, the bootstrap coefficient pvalue is the proportion of the bootstrap estimated coefficients that are greater than zero.

Table 14 and Table 15 report the initial coefficient estimate (Coef) and pvalue from Table 10 and Table 11, along with the bootstrap coefficient estimate (Boot. Coef), the bootstrap pvalue, and the 90% Confidence Interval (CI) limits for the coefficient estimate from the bootstrap resampling. The CI limits correspond to the fifth and ninety-fifth quantiles of the bootstrap samples; the estimated bootstrap coefficient is simply the mean value of all the bootstrap samples that fall within the 90% CI.

All of the initial coefficient estimates fall within the bootstrap 90% CI. With one exception, the bootstrap validation returns p-values that are consistent with (or better than) the initial model p-values. The exception is in the Latent Bias Model: the p-value for the factor loading between *Bias* and *Worried* (λ_4) increases from 0.164 to 0.392 in the bootstrap. The increased pvalue indicates that perhaps *Worried* should not be a predictor for *Bias.* Removing *Worried* as a factor for *Bias* is consistent with the Bias Path Analysis Model, which retains *Worried* as a predictor for *Dif* but not as a predictor for *Trip*.

Table 14: Bootstrap resampling in the Latent Bias Model

Latent Bias Model		Coef.	Boot. Coef	Pvalue	Boot. Pvalue	Boot. 90% CI – lower bound	Boot. 90% CI – upper bound

Latent Bias Model		**Coef.**	**Boot. Coef**	**Pvalue**	**Boot. Pvalue**	**Boot. 90% CI – lower bound**	**Boot. 90% CI – upper bound**
Bias=~ Unreal	Λ_1	1.0	0.1	–	–	–	–
Bias=~ Prepared	Λ_2	−0.378	−0.519	0.086	0.034	−1.643	−0.02
Bias=~ Worried	$\mathbf{\Lambda_5}$	**0.173**	**0.032**	**0.164**	**0.392**	**−0.637**	**0.465**
Bias~ Graded	Λ_2	−1.296	−1.336	0.016	0.006	−2.563	−0.448
Bias~ Yr2015	Λ_4	−0.853	−0.906	0.227	0.078	−2.157	0.213
Bias~ Team	Λ_5	−5.216	−5.098	0.023	0.028	−9.526	−0.564
Trip~ Bias	B_1	0.05	0.057	0.141	0.072	−0.007	0.157
Trip~ Yr2015	Γ_{14}	−0.467	−0.464	0.003	0.014	−0.702	−0.196
Trip~ Team	Γ_{15}	−0.852	−0.782	0.037	0.062	−1.532	0.033
Trip~ Quiz	Γ_{18}	−2.417	−2.329	0.022	0.002	−3.675	−1.063
Dif~ Trt1	Γ_{21}	0.258	0.23	0.069	0.111	−0.093	0.495
Dif~ Base Stress	Γ_{26}	0.108	0.101	0.244	0.109	−0.035	0.238
Dif~ Evaluation	Γ_{27}	−0.668	−0.615	0.07	0.044	−1.131	−0.043

Table 15: Bootstrap resampling in the Bias Path Analysis Model

Bias Path Analysis Model		**Coef.**	**Boot. Coef**	**Pvalue**	**Boot. Pvalue**	**Boot. 90% CI – lower bound**	**Boot. 90% CI – upper bound**
Unreal~ Graded	α_{12}	−1.12	−1.261	0.046	0.005	−2.584	−0.373
Unreal~ Yr2015	α_{14}	−0.933	−1.007	0.238	0.077	−2.33	0.126
Unreal~ Team	α_{15}	−4.805	−5.851	0.03	0.01	−10.581	−2.347
Likely.Against~ Fami	α_{43}	−0.119	−0.127	0.233	0.067	−0.283	0.016
Likely.Against~ Yr2015	α_{44}	−0.172	−0.176	0.042	0.019	−0.319	−0.031
Likely.Against~ Team	α_{45}	−0.513	−0.477	0.07	0.009	−0.819	−0.123
Worried~ Team	α_{55}	−2.631	−3.665	0.225	0.099	−11.139	1.379
Worried~ Base Stress	α_{56}	0.532	0.998	0.176	0.081	−0.134	4.178
Trip ~ Unreal	β_{11}	0.071	0.064	0.056	0.059	−0.001	0.13
Trip~ Likely.Against	β_{14}	−0.493	−0.507	0.116	0.07	−1.209	0.058
Trip~ Yr2015	γ_{14}	−0.546	−0.554	0.001	0.001	−0.866	−0.275
Trip~ Team	γ_{15}	−1.035	−0.99	0.015	0.023	−1.718	−0.28
Trip~ Quiz	γ_{18}	−2.415	−2.373	0.024	0.004	−3.672	−1.075
Dif~ Likely.Against	β_{24}	0.421	0.399	0.103	0.219	−0.785	1.21
Dif~ Worried	β_{25}	−0.067	−0.051	0.195	0.197	−0.148	0.052
Dif~ Trt1	γ_{21}	0.223	0.174	0.158	0.205	−0.182	0.489
Dif~ BaseStress	γ_{26}	0.147	0.141	0.107	0.072	−0.025	0.321
Dif~ Evaluation	γ_{27}	−0.682	−0.61	0.064	0.059	−1.128	0.035

With the assurance that the bootstrap resampling supports the validity of the Bias Path Analysis Model and, for the most part, the Latent Bias Model, we turn to the RMSEA fit measure.

3.3.2 Root Mean Square Error of Approximation

The RMSEA value—represented by ε in the equations below—is calculated from F_0, the final value of the estimator used in the model fitting algorithm (the algorithm iterates through models until the fitting function, F, reaches a minimum value, F_0). F is expected to follow a noncentral Chi-squared distribution, specified by two parameters: the degrees of freedom d[7] and the noncentrality parameter λ. The model Chi-squared value, T, is the expected value of the distribution, and is calculated as a function of F_0 and the sample size, n: $T = E[\chi^2_{d,\lambda}] = (n-1)F_0$. The RMSEA value is defined as:

$$\varepsilon = \sqrt{\frac{T-d}{nd}} \tag{5}$$

The relationship between F_0, T and ε is explained in greater detail in [25]. General consensus is that RMSEA less than 0.05 represents a good fit, and RMSEA greater than 0.08 is a poor fit [23]. Because RMSEA follows a known distribution, researchers can estimate confidence intervals to test the significance and power of hypothesis tests of model fit. Following the approach outlined in [25], we use the test of "not-close fit" to determine whether we should reject a proposed model.

The null hypothesis in the test of not-close fit is that the RMSEA value is greater 0.05, the upper limit for a close-fitting model. The alternative hypothesis is that the model is a close fit to the data:

$$\begin{cases} H_0: \varepsilon > 0.05 \\ H_a: \varepsilon \leq 0.05 \end{cases} \tag{6}$$

To test this hypothesis, we establish a confidence interval (CI) around the estimated RMSEA value. There are three possible outcomes:

- **Not-Close Fit:** Do not reject the null hypothesis if the lower limit of the CI is greater than 0.05.
- **Close Fit:** Reject the null hypothesis that the model is not a close fit if the upper limit of the CI is less than 0.05.
- **Indeterminate:** The test is indeterminate if the CI includes 0.05.

For both models, the 90% RMSEA CI estimated by *lavaan* is [0,0.0001], meaning we can reject the hypothesis of not-close fit and conclude that our models are a close fit to the data. Our objective in the next section is to determine the sample size needed for a high confidence in this result.

3.3.3 The Power of the Test

We must now examine the power of that result—that is, the probability that the models are actually a close-fit for the data.

[7] $d \sim$number of observed covariances (i.e. $p(p+1)/2$ where $p \sim$ number of observed variables) minus the number of parameters estimated in the model [19].

To do this we specify a new hypothesis, again following the approach suggested in [25]. The null hypothesis is that the RMSEA value is 0.05, and the alternative hypothesis is that the RMSEA value is a very close fit, $\varepsilon = 0.01$:

$$\begin{cases} H_0 : \varepsilon = 0.05 \\ H_a : \varepsilon = 0.01 \end{cases} \tag{7}$$

We calculate the probability that a model with $\varepsilon = 0.01$ will not be recognized as a close-fit to the data. We do this by exploiting the relationship between the RMSEA value and the Chi-squared distribution. By definition [26], the expected value of a non-central Chi-squared distribution is $E[\chi^2_{d,\lambda}] = d + \lambda$. This allows us to calculate the Chi-squared non-centrality parameter, λ, as a function of the RMSEA:

$$\lambda = \varepsilon^2 n d \tag{8}$$

With $d = 30$ (selected because $d = 31$ for both the Latent Bias Model and the Bias Path Analysis Model), we use $\varepsilon_0 = 0.05$ and $\varepsilon_a = 0.01$ to calculate $\lambda_0 = \varepsilon_0^2 n d$ and $\lambda_a = \varepsilon_a^2 n d$. The associated Chi-squared distributions are plotted in Figure 9. The solid line is the distribution associated with the null hypothesis, and the dotted line is the alternative hypothesis. Figure 9 compares three sample sizes: (a) *n = 41*, (b) *n = 204,* and (c) *n = 338*.

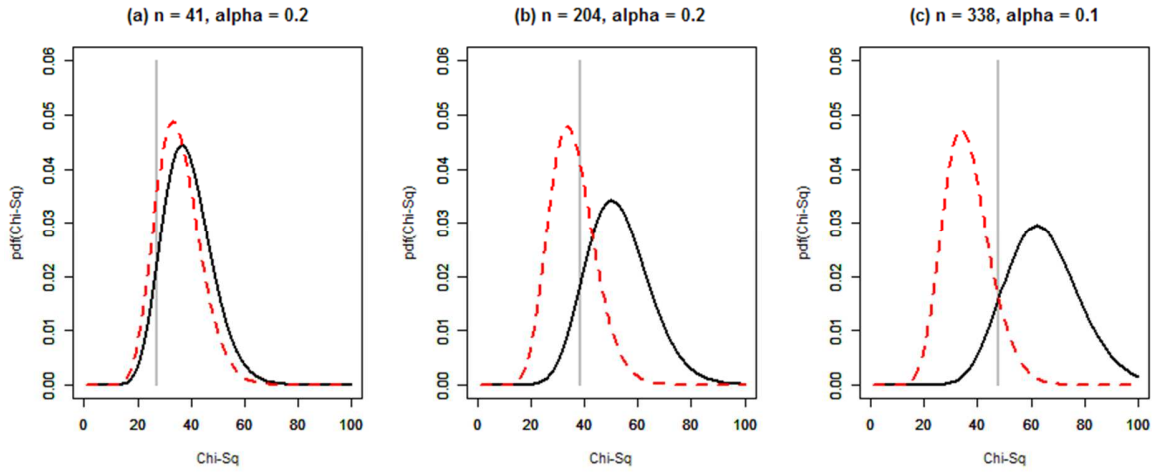


Figure 9: Power analysis and sample size for a model with 30 degrees of freedom. (a) shows the OSU NPP Simulator Experiment sample size, (b) is the recommended sample size for $\alpha = 0.2, \pi = 0.8$, and (c) is the recommended sample size for $\alpha = 0.1, \pi = 0.9$.

The thin gray line in Figure 9 marks the critical point, χ^2_c: we do not reject the null hypothesis if the Chi-squared value is greater than χ^2_c. The critical point is calculated from λ_0 according to Equation 8:

$$\Pr(\chi^2_{d,\lambda_0} < \chi^2_c) = \frac{\alpha}{2} \tag{9}$$

The power of the test, π, is the probability that we do not accept H_0 when H_a is true. In Figure 9, this corresponds to the area under the dotted red line to the left of the critical point:

$$\pi = Pr(\chi^2_{d,\lambda_a} < \chi^2_c) \tag{10}$$

Figure 9 (a) illustrates the difficulty in differentiating between a close-fit and a not-close fit with a small sample size: with *n = 41*, $\pi_{41} = 0.16$. Figure 9 (b) and (c) show how the power improves as sample size increases; with *n = 204* and a critical point defined by $\alpha = 0.2$, the discriminating power is 80%. Increasing *n* to 338 yields a power of 90% with a more stringent $\alpha = 0.1$. We selected the sample sizes *n = 204 and n = 338* by setting α and substituting different values of *n* into λ_0 and λ_a (Equation 8) until Equations 9 and 10 were true for the desired power level.

In summary, a sample of *n = 300* is recommended to obtain conclusive results to the test of not-close fit with a power of 80% for both the Latent Bias Model and the Bias Path Analysis Model. Note that the recommended *n* will vary with the degrees of freedom in the model. The sample size for the Latent Bias Model and Bias Path Analysis Model is high because each model has only 31 degrees of freedom.

3.3.4 Sample size requirements for testing mediated effects

In general, *n = 300* will yield high confidence in the estimated model parameters. The exception is mediated effects, which are more difficult to estimate. To estimate the sample size needed to have high confidence in the mediated effects, we use the guidelines proposed in [27].

Table 1: Cohen's effect size thresholds [27].

Mediated Effect	***Standardized Coefficient***
Small (S)	0.18
Halfway (H)	0.26
Medium(M)	0.39
Large (L)	0.59

The sample size depends on the strength of the mediated effect of interest—in this case, the magnitude of the standardized coefficient. Cohen [28] proposed widely-accepted thresholds for categorizing effects as small (S), medium (M) or large (L) in regression analysis; Fritz and MacKinnon [27] add a halfway (H) threshold that evaluates effects halfway between small and medium. The recommended standardized coefficient thresholds corresponding to small, halfway, medium and large effects are listed in Table 16. Using Monte Carlo simulation, Fritz and MacKinnon estimate the minimum sample size needed to detect each effect size in mediated models. Typically, the sample size required to detect mediated effects is much greater than sample sizes needed to assess direct effects. Table 17 lists estimates of the minimum sample size needed to detect mediated effects using the percentile bootstrap validation approach with a power of 0.8.[8]

Table 17: Recommended minimum sample size required for bootstrap power analysis ($\pi = 0.8$). Effects are listed as distal-proximate effects, i.e. SM indicates Small distal effect (a in Figure 5) and a Medium proximal effect (b in Figure 5). S, M and L correspond to Cohen's criteria for Small, Medium, and Large Effects; H represents Halfway between S and M. From Table 3 in [27].

Mediated effects	*SS*	*SH*	*SM*	*SL*	*HS*	*HH*	*HM*	*HL*	*MS*	*MH*	*MM*	*ML*	*LS*	*LH*	*LM*	*LL*
n	558	412	406	398	414	162	126	122	404	124	78	59	401	123	59	36

Based on Table 17, a sample of $n = 41$ should capture significant Large-Large mediated effects, that is, effects that have a large effect on bias when bias has a large effect on the response variable. With lower power, our small sample should detect significant Large-Medium and Medium-Large mediated effects.

[8] Fritz and MacKinnon review six different tests for the significance of the effect. We selected the percentile bootstrap one of the most powerful and straightforward tests, but the other approaches might be preferable in other contexts.

Smaller effects (SS, SL, etc.) may be found to be significant, but power associated with these estimates will be low.

3.4 Recommended Sample Size for Future Research

Figure 10 diagrams the structure of a model that might be used in future experiments. The proposed model is based on the Bias Causal Model (Figure 1), and includes three latent biases, each measured by three indicators.

We retain the expectation from the Bias Causal Model that *Incentive* acts on *Cavalier Behavior*, but we separate the effects of *Cavalier Behavior* and *Prominent Hypothesis* on the response variables *Trip* and *Dif*. This corresponds to the results from this analysis that show *Trip* is a function of *Unreal* and *Likely.Dist,* while *Dif* depends on *Likely.Dist* and *Worried*. These results indicate that reducing "Bias" to a single effect is too simplistic, and a more nuanced model is required.

The heavy lines in Figure 10 correspond to the model we would expect based on the Bias Path Analysis Model, with *Unreal* corresponding to *Cavalier Behavior* and *Likely.Dist* to *Prominent Hypothesis*. The other possible links are shown by the lighter lines in Figure 10.

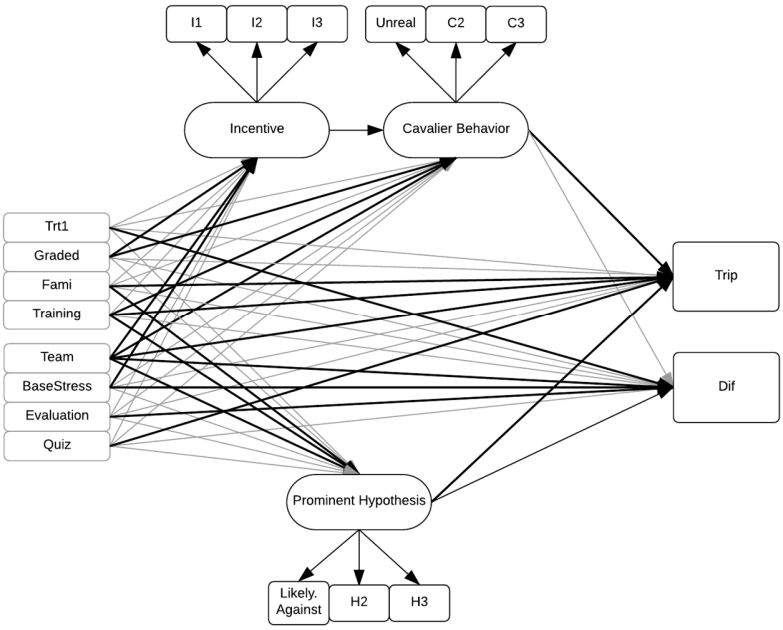
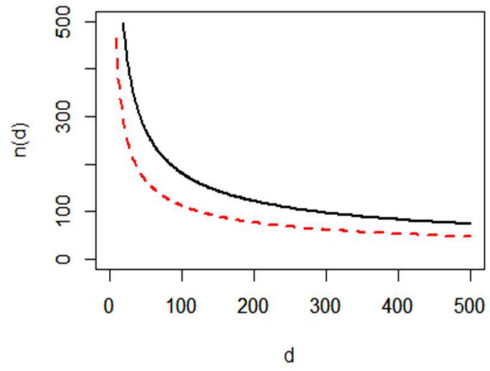


Figure 10: Sample Bias SEM for future work.

As illustrated in Section 3.3.4, the recommended sample size varies with the degrees of freedom in the model. If we retain all possible links in the model, *d = 82*. If we include only the links suggested by the Bias Path Analysis Model (i.e. the darker lines in Figure 10), *d = 106*. In general, removing links between variables increases *d*; removing variables from the model or adding links between variables decreases *d*. As Figure 11 shows, the recommended sample size stabilizes as degrees of freedom in the model

increase. In Figure 11, the black line corresponds to $\alpha = 0.1, \pi = 0.9$, and the dotted line to the less stringent criteria $\alpha = 0.2, \pi = 0.8$. For *d = 100*, the minimum recommended *n* ranges from 89 to 139.



Recommended Sample Size:		
d	$\boldsymbol{\alpha = 0.1, \pi = 0.9}$	$\boldsymbol{\alpha = 0.2, \pi = 0.8}$
30	372	223
50	272	166
82	203	125
106	175	109
150	143	90
200	122	77
30	372	223

Figure 11: Recommended sample size as a function of degrees of freedom. LEFT: The solid line shows n(d) for $\alpha = 0.1, \pi = 0.9$; the dotted line corresponds to $\alpha = 0.2, \pi = 0.8$. RIGHT: recommended sample size for the anticipated degrees of freedom in the proposed Bias SEM.

Recall from Figure 9 that the recommended sample size for testing the Bias Path Analysis Model is *n = 300,* more than twice the recommended *n* for the model proposed in Figure 10. This is because *d = 100* in the proposed model, and *d = 28* in the Latent Bias Model and *d = 30* in the Bias Path Analysis Model. In the proposed model, with three indicators for each latent variable, the data collected in each scenario will provide more information than the per-scenario data collected in the OSU NPP Simulator Experiment. In essence, the proposed model allows us to extract more information per sample rather than increasing the number of samples required.

3.4.1 Alternative approach for estimating sample size

Instead of relying on statistical tests to estimate sample size, many researchers use Monte Carlo simulations to predict how their model will behave for various sample sizes. Based on a survey of Monte Carlo simulations of SEM research, Westland [29] recommends choosing *n* based on *r*, the ratio of indicators to latent variables, according to Equation 11:

$$n \geq 50\, r^2 - 450\, r + 1100 \tag{11}$$

In Figure 10, each latent variable has three associated indicators. For *r=3*, Westland's recommended sample size is *n=200*. At *n=200*, the power calculated using the test of not-close fit [25] with $\alpha = 0.1$ is $\pi = 0.75$ for models with *d = 50* and $\pi = 0.94$ for models with *d = 100*.

3.4.2 Recommended Sample Size for Mediated Effects

Recall from Table 17 that detecting mediated effects may require larger samples. Researchers should use the findings from [27] in conjunction with the RMSEA test of not-close fit or Westland's Monte Carlo-based guideline to design experiments to test mediated effects.

4 DISCUSSION

This research proposes a method to capture the impact of simulator bias on data collected in simulator facilities. This approach allows us to study human reliability in a simulator environment, then calibrate the data to better represent operator actions in a control room.

The examples below use the Bias Path Analysis Model proposed in Section 3 to quantify the effects of the simulator environment. In Section 4.1 we quantify the expected change in *Trip* time when we shift from ungraded to graded scenarios. In Section 4.2 we use the *Trip* time observed in the simulator scenario to estimate an expected *Trip* time in an actual NPP control room. These examples illustrate how this method of modeling simulator bias could be applied in HRA research.

Note that we do not expect the numerical results in these examples to represent experienced operators—remember that this model is based on a small sample of student operators—but we provide these examples to illustrate the potential applications of the method.

4.1 QUANTITATIVE ANALYSIS OF THE VARIATIONS IN SIMULATOR BIAS

In this example, we examine how experiment phase (*Graded*) impacts response times in the Bias Path Analysis Model.

Let **Condition 0** represent the nominal scenario, i.e. $Graded_0 = 0$, and **Condition 1** represent the added incentive, i.e. $Graded_1 = 1$. In the Bias Path Analysis Model, the relationship between *Graded* and *Trip* is indirect; a change in *Graded* impacts *Unreal*, and the change in *Unreal* impacts *Trip*. From the model, we estimate the change in *Unreal*, $\Delta Unreal$, as a function of the change in Graded, $\Delta Graded$:

$$\Delta Unreal = \alpha_{12} \Delta Graded \tag{12}$$

We use this relationship to estimate the change in Trip, $\Delta Trip$:

$$\Delta Trip = \beta_{11} \Delta Unreal = \beta_{11} \alpha_{12} \Delta Graded \tag{13}$$

From Table 11, $\alpha_{12} = -1.12$ is the Bias Path Analysis Model coefficient for the effect of *Graded* on *Unreal*, and $\beta_{11} = 0.071$ is the effect of *Unreal* on *Trip*. Inserting the values from Table 11 yields

$$\Delta Trip = -0.08$$

Recall that *Trip* and *Dif* are scenario normalized response times, that is, the measured time divided by the mean time for that scenario. This means, holding all other factors constant, we expect an eight percent decrease in *Trip* in the *Graded* phase of the experiment.

4.2 QUANTITATIVE ESTIMATE OF OVERALL BIAS EFFECTS

In this example we demonstrate how data collected in a simulator could be used to estimate response times in a NPP control room. To do this, we estimate the expected "real world" values for the Bias Measures. The difference between the expected control room values and the average simulator values is the expect change in the Bias Measures that would result from moving from the simulator to the control room. The change in the Bias Measures is inserted into the Bias Path Analysis Model to predict the corresponding change in *Trip* and *Dif* (denoted by $\Delta Trip,\ \Delta Dif$).

Let **Condition 0** represent in the simulator, and the hypothesized **Condition 1** represent the actual control room. The Bias Measures in the Bias Path Analysis Model are *Unreal, Likely.Against,* and *Worried.* We estimate the change in the Bias Measures when we move from the simulator to the control room, then adjust *Trip* and *Dif* accordingly.

From Table 11, β_{11} and β_{14} are *Trip*'s coefficients for *Unreal* and *Likely.Against* in the Bias Path Analysis Model; β_{24} and β_{25} are *Dif*'s coefficients for *Likely.Against* and *Worried*. The expected change in *Trip* and *Dif* resulting from a change in the Bias Measures is therefore:

$$\begin{cases} \Delta Trip = \beta_{11}\Delta Unreal + \beta_{14}\Delta Likely.Against \\ \Delta Dif = \beta_{24}\Delta Likely.Against + \beta_{25}\Delta Worried \end{cases} \quad (14)$$

Table 18 explains how we estimate the expected change in the Bias Measure based if we were to shift from the simulator to the control room. We use data observed in the simulator to estimate expected simulator values (Condition 0), and propose reasonable estimates for conditions in the actual NPP control room (Condition 1).

Table 18: Expected values of Bias Measures in a control room (rather than a simulator).

Expected change in Bias Measures moving from the simulator (Condition 0) to the control room. and estimated values in the control room (Condition 1)
Unreal: $\Delta Unreal = Unreal_1 - Unreal_0$, where the variable $Unreal_1$ represents the operator feeling of unreality in the control room (Condition 1) and $Unreal_0$ is the feeling of unreality in the simulator (Condition 0). *Unreal* has five levels; the mean reported *Unreal* in this experiment is Level 3. We assume the control room value $Unreal_1 = 1$, representing the fact that the feeling of unreality is at a minimum in a real control room. We set $Unreal_0 = 3$, the mean *Unreal* value in the simulator. The change is therefore a decrease of 2 levels: $\Delta Unreal = Unreal_1 - Unreal_0 = 1 - 3 = -2$ This measure can be improved by surveying operators who have responded to design basis accidents while on shift. Preferably, the reality of the simulator vs control room environments will be assessed through multiple survey questions.
Prepared: Although *Prepared* is not in the final model, for the sake of completeness we review how to estimate *Prepared* in the control room. Operators in the control room are not expecting an accident to occur, but their extensive training should make them reasonably well prepared when an accident happens. We therefore estimate that $Prepared_1 = Prepared_0$, assuming that operators who are equipped to respond to a given scenario in the simulator are equally well prepared to respond to the same accident in the control room. Therefore, $\Delta Prepared = Prepared_1 - Prepared_0 = 0$ Alternatively, $Prepared_1$ could be quantified as a function of the time spent training on the accident and recency of the training.
Likely: *Likely* is a conditional probability; when an accident occurs, *Likely* is the operator's assessment of the probability that a specific set of events is the root of the accident. *Likely.For* is the expectation of the event that actually occurs; *Likely.Against* is the expectation of accidents that do no actually occur. In the Bias Path Analysis Model, *Trip* and *Dif* are functions of *Likely.Against* (*Likely.For* is dropped from the model).

Theoretically, Prominent Hypothesis Bias in the control room is zero, so we set $Likely.Against_1 = 0$. The average *Likely.Against* in the simulator is $Likely.Against_0 = 0.14$. This yields:

$$\Delta Likely.Against = Likely.Against_1 - Likely.Against_0 = 0 - 0.14 = -0.14$$

Alternatively, it is possible that operators bring a hypothesis bias into the control room, based on their training and experience. We can estimate $Likely.For_1$ and $Likely.Against_1$ by surveying operators about their expectations in conjunction with reviewing training and plant history.

Worried: *Worried* is 3-level ordinal variable. The student operators report whether or not they are worried that their performance will impact their course grade. Although this is not an exact parallel, we set $Worried_1 = 3$, the maximum value on *Worried*'s 3-level scale. The mean level for *Worried* in the experiment is 2, resulting in a one-level increase in *Worried* from simulator to control room:

$$\Delta Worried = Worried_1 - Worried_0 = 3 - 2 = 1$$

Inserting the appropriate values from Table 11 into Equation 14 and solving for $Trip_1$ and Dif_1 rather than $\Delta Trip$ or ΔDif results in

$$\begin{cases} Trip_1 = Trip_0 + 0.071 \times (-2) - 0.493 \times (-0.14) = Trip_0 - 0.073 \\ Dif_1 = Dif_0 + 0\ .421 \times (-0.14) - 0.067 \times (1) = Dif_0 - 0.126 \end{cases} \quad (15)$$

If, as in the previous section, the average time to trip the reactor is 530s, and if the time to exit E0 is an additional 895s, the time to trip and then to diagnose in a control room ($SecTrip_1$ and Sec_Dif_1) can be estimated as

$$\begin{cases} SecTrip_1 = SecTrip_0 \times (1 - 0.072) = 0.927 \times 530s = 491s \\ SecDif_1 = SecDif_0 \times (1 - 0.126) = 0.874 \times 895s = 782s \end{cases} \quad (16)$$

This corresponds to tripping the reactor about half a minute faster in the control room, and exiting E0 approximately two minutes faster in the control room. A total reduction in scenario time of 2.5 minutes is equivalent to approximately ten percent of the time required in the simulator.

A few caveats: First, the significance of a ten percent reduction in time depends on the scenario and the time available. In some instances, this time difference is trivial; in other scenarios, these results might mean the difference between failure in the simulator but success in the control room.

Second, as always, this is a demonstration experiment, meant to illustrate the possibilities of our methods. We do not expect these differences in times between the simulator and the control room to be valid for experienced operators. We understand that, in general, operators are expected to require more time in the control room because of prominent hypothesis effects, etc.

5 CONCLUSION

This paper introduced a framework for addressing simulator bias. Biases are categorized as either Motivational Biases or Environmental Biases. Environmental Biases are the effects of any discrepancies between the physical environment of the control room and the simulator; Motivational Biases refer to the cognitive impact of the artificial nature of a simulator exercise. The biases defined in this research are presented as a starting point for studying simulator bias and considering bias in other applications of

simulator research. We expect that the simulator biases will be expanded and refined as this area of research grows.

We conducted an experiment with student operators to demonstrate how simulator bias can be manipulated. Using SEM, we developed two models of simulator bias: the Latent Bias Model, which estimates a quantitative value for simulator bias; and the Simulator Bias Model, which estimates the relationships between the scenario conditions, the bias measures, and the response variables. These models illustrate two ways to use SEM to analyze simulator bias.

Using these models, we demonstrated how to assess model fit and estimate the requisite sample size to build a statistically significant model. Finally, we illustrate how the bias models can estimate operator performance data in a functioning control room using data collected in a simulator environment. Although the data are obtained from student operators and are not expected to fully represent expert operator behavior, the methods and analysis illustrated in this research introduce a formal approach to addressing bias in simulator studies.

Our research demonstrates that the effects of the simulator environment—abstract and intangible though they may be—are not beyond the reach of experimental science. With proper instruments, SEM can help assess the invisible factors that influence our research.

5.1 Recommendations for Future Research

Our results suggest that observing approximately 200 scenarios will yield models of sufficient power to make conclusive judgments about some of the simulator effects observed in this experiment. SACADA is expected to collect information from 560 scenarios per year [1], more than twice the sample size needed for robust models of simulator bias. Table 19 lists recommendations for designing experiments to measure simulator bias or to consider simulator bias in other simulator-based experiments.

Table 19: Recommendations for considering simulator bias in experiment design

Recommendations for Simulator Bias Experiments
Develop multiple measures for each expected Bias. Multiple measures of each bias should be developed, preferably a mix of independent, subjective and objective measures, perhaps assessed by the operator and by an independent observer. When possible, more informative measures should be developed; yes or no questions should be replaced with scaled response questions.
Do not assume Simulator Bias can be treated as a single variable. In the Path Analysis Model, *Unreal* and *Likely.Dist* affect *Trip*, while *Likely.Dist* and *Worried* affect *Dif*. This is because the effect of the *Prominent Hypothesis* does not act in concert with *Cavalier Behavior*. Multiple measures for each bias allow for models with multiple, interacting latent variables.
Control crew characteristics in the experimental design. We did not control for team dynamics (*Team*), operator tendencies (*BaseStress*) or operator expertise (*Quiz, Evaluation*) in the OSU NPP Simulator Experiments, but these factors have a significant impact on bias and response times. Rather than randomly assigning crews, future research should control for crew characteristics in the experimental design.
Include specific response variables in the experimental design. For simplicity, we limited our analysis to two response variables (*Trip* and *Dif*), but other response variables of interest might include number of diagnoses considered before identifying the accident, reactor operator confidence throughout the scenario, etc. Measures of operator stress will be important in experiments related to

HRA models such as SPAR-H.
Develop dynamic measures of simulator bias. Ideally, future studies will propose a method to evaluate biases periodically throughout the scenario rather than as static factors at the end.
Collect data from null scenarios and accident scenarios. We can reduce the gap between simulator and control room conditions by including simulator sessions that focus on routine work.
Use experienced operators. Our student operator models demonstrate the feasibility and potential of this method, but reliable models of simulator bias require data from a reliable source.

The biggest challenge to this effort is not scope or scale, but in developing data collection instruments to measure bias effects. Multiple, independent measures of each bias are necessary to build robust models that capture the different simulator biases and their effects. The surveys we developed for this study are a starting point for this effort.

We do not attempt to prove or disprove our theory with this demonstration experiment; we merely suggest that we can make concrete hypotheses about the impact of the simulator environment, and the evidence to support or reject these hypotheses may be only two hundred simulator scenarios away.

ACKNOWLEDGEMENTS

We would like to thank Atul Gupta, Nan Wang, Jordan Jurgensmier and the students in the OSU NPP Systems & Operations Course.

This research is being performed using funding received from the DOE Office of Nuclear Energy's Nuclear Energy University Programs, the Nuclear Regulatory Commission (NRC) Graduate Fellowship Program (NRC-HR-13-G-38-0047), and an NRC educational grant (NRC-38-10-962). The OSU NPP Simulator Facility is funded in part by the DOE (DE-NE0000450).

6 REFERENCES

[1] Y. J. Chang, D. Bley, L. Criscione, B. Kirwan, A. Mosleh, T. Madary, R. Nowell, R. Richards, E. M. Roth, S. Sieben and A. Zoulis, "The SCADA Database for Human Reliability and Human Performance," *Reliability Engineering & System Safety,* vol. 125, no. Special Issue, pp. 117-133, 2014.

[2] A. Gupta, Development of Boiling Water Reactor Nuclear Power Plant Simulator for Human Reliability Analysis Education and Research, Columbus, OH: Ohio State University Master's Thesis, 2013.

[3] "Simulator Studies for HRA Purposes," OECD NEA CSNI, Budapest, Hungary, 2009.

[4] R. B. Shirley, C. S. Smidts, N. Wang and A. Gupta, "Elimination of bias in NPP control room simulator experiments," in *ANS Winter Meeting*, Anaheim, CA, 2014.

[5] M. A. Hotchkiss, C. Biddle and M. Fallacaro, "Assessing the authenticity of the human simulation experience in anesthesiology.," *AANA Journal,* vol. 70, no. 6, pp. 470-3, 2002.

[6] M. Levine, "Hypothesis theory and non-learning despite ideal SR-Reinforcement Contingencies," *Psychological Review,* vol. 78, no. 2, pp. 130-40, 1971.

[7] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science,* vol. 185, no. 4157, pp. 1124-1131, 1974.

[8] D. L. Paulhaus, "Measurement and control of response bias," in *Measures of personality and social psychological attitudes*, vol. 1, San Diego, Academic Press, 1991, pp. 17-59.

[9] M. Rubin, "The Perceived Awareness of the Research Hypothesis Scale: Assessing the influence of demand characteristics.," 28 12 2016. [Online]. Available: https://doi.org/10.6084/m9.figshare.4315778.v2. [Accessed 21 10 2017].

[10] J. Polak and P. Jones, "The acquisition of pre-trip information: A stated preference approach," *Transportaton,* vol. 20, pp. 179-98, 1993.

[11] H. Koutsopoulos and A. Polydoropoulou, "Travel simulators for data collection on driver ehavior in the presence of information.," *Transportation Research: Part C,* vol. 3, no. 3, pp. 143-59, 1995.

[12] J. G. Casali and W. W. Wierwille, "The effects of various design alternative on moving-based driving simulator discomfort.," *Journal of Sound and Vibration,* vol. 90, no. 4, pp. 610-1, 1980.

[13] G. E. Prussia, K. A. Brown and P. G. Willis, "Mental models of safety: do managers and employees see eye to eye?," *Journal of Safety Research,* vol. 34, pp. 143-156, 2003.

[14] Y. Chang and A. Mosleh, "Cognitive Modeling and Dynamic Probabilistic Simulation of Operator Crew Response to Complex System Accidents, Parts 1-5," *Reliability Engineering & System Safety,* vol. 92, no. 8, pp. 997-1013, 2007.

[15] A. Spurgin and G. Petkov, "Advanced Simulator Data Mining for Operators' Performance Assessment," *Studies in Computational Intelligence,* vol. 5, pp. 487-514, 2005.

[16] R. B. Shirley, C. Smidts, R. Boring, A. Mosleh and Y. Li, "Science-Based HRA: Experimental Comparison of Operator Performance to IDAC Simulations," in *9th International Conference on Nuclear Plant Instrumentation, Control & Human-Machine Interface Technologies*, Charlotte, NC, 2015.

[17] N. Wang, R. Benish and C. Smidts, "A preliminary analysis on Performance Shaping Factors in a Science-Based HRA validation method," in *ANS Student Conference*, State College, PA, 2014.

[18] A. Skrondal and S. Rabe-Hesketh, "Structural Equation Modeling: Categorical Models," Wiley, 2005. [Online]. Available: http://www.gllamm.org/SEMcat.pdf. [Accessed 10 6 2016].

[19] T. Raykov and G. A. Marcoulides, A First Course in Structural Equation Modeling, Mahway, NJ:

Lawrence Erlbaum Associates, Inc., 2000.

[20] Y. Rosseel, "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software,* vol. 48, no. 2, pp. 1-36, 2012.

[21] D. A. Kenny, "Mediation," 22 5 2016. [Online]. Available: http://www.davidakenny.net/cm/mediate.htm. [Accessed 17 6 2016].

[22] D. P. MacKinnon, A. J. Fairchild and M. S. Fritz, "Mediation Analysis," *Annu. Rev.Psuchol.,* vol. 58, pp. 593-614, 2007.

[23] D. Hooper, J. Coughlan and M. Mullen, "Structural Equation Modeling: Guidelines for Determining Model Fit," *Electronic Journal of Business Research Methods,* vol. 6, no. 1, pp. 53-60, 2008.

[24] M. Kutner, C. J. Nachtsheim and J. Neter, in *Remedial Measures for Evaluating Precision in Nonstandard Situations--Bootstrapping*, Singapore, McGraw-Hill Education, 2008, pp. 458-464.

[25] R. C. MacCallum, M. W. Browne and H. M. Sugawara, "Power Analysis and Determination of Sample Size for Covariance Structure Modeling," *Psychological Methods,* vol. v, no. 2, pp. 130-149, 1996.

[26] K. Krishnamoorthy, "Noncentral Chi-square Distribution," in *Handobok of Statistical Distributions with Applications*, Chapman and Hall/CRC, 2015, p. 259.

[27] M. S. Fritz and D. P. MacKinnon, "Required Sample Size to Detect the Mediated Effect," *Psychol Sci.,* vol. 18, no. 3, pp. 233-239, 2007.

[28] J. Cohen, Statistical power analysis for the behavioral sciences, HIllsdale, NJ: Erlbaum, 1988.

[29] J. C. Westland, "Lower bounds on sample size in structural equation modeling," *Electronic Commerce Research and Applications,* vol. 9, pp. 476-487, 2010.

[30] R. Benish, C. Smidts and R. Boring, "A pilot experiment for science-based Human Reliability Analysis (HRA) validation," in *Probabilistic Safety Assessment 2013*, Columbia, SC, 2013.