

QUANTIFYING QUALITY: USING QUANTITATIVE METHODS TO EVALUATION
OBSERVATION QUALITY AND ITS IMPACT ON INCIDENT REDUCTION

A Thesis
by
CHARLES RIGGS MATTHEWS

Submitted to the School of Graduate Studies
at Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF ARTS

May 2022
Department of Psychology

QUANTIFYING QUALITY: USING QUANTITATIVE METHODS TO EVALUATE
OBSERVATION QUALITY AND ITS IMPACT ON INCIDENT REDUCTION

A Thesis
by
CHARLES RIGGS MATTHEWS
May 2022

APPROVED BY:

–
Yalçın Açıkgöz, Ph.D.
Chairperson, Thesis Committee

–
Shawn Bergman, Ph.D.
Member, Thesis Committee

–
Timothy Ludwig, Ph.D.
Member, Thesis Committee

–
Rose Mary Webb, Ph.D.
Chairperson, Department of Psychology

–
Marie Hoepfl, Ed.D.
Interim Dean, Cratis D. Williams School of Graduate Studies

Copyright by Charles Riggs Matthews 2022
All Rights Reserved

Abstract

QUANTIFYING QUALITY: USING QUANTITATIVE METHODS TO EVALUATION OBSERVATION QUALITY AND ITS IMPACT ON INCIDENT REDUCTION

Charles Riggs Matthews
B.S. Clemson University
M.A. Appalachian State University

Chairperson: Yalçın Açıköz, Ph.D.

Behavior-based safety (BBS) management systems are effective occupational safety programs that rely on behavioral data collection, intervention, and evaluation to assess intervention effectiveness. This data is collected on observation checklists via peer-to-peer behavioral observations. While there has been research confirming the effectiveness of observation reports on incident prevention, there has been limited research on how checklist quality moderates observation checklist's impact on incident reduction. This may be important, as the BBS iterative process may yield inconsistent results if the data collected is inaccurate. The current study investigates checklist design and response, operationalizing quality checklist design as having more discrimination items, more safety-confirmative items, and more free-response questions. It operationalizes response quality as fixed-response variation and free-response length. All operationalizations of checklist design moderated the effectiveness of observation checklists on interventions such that observation checklists with greater quality were more effective at reducing incident likelihood. Both operationalizations of response quality had insignificant effects on observation effectiveness.

Keywords: safety analytics; observation quality; checklist quality; behavior based safety

Acknowledgements

I would like to express my gratitude to my committee chair, Yalcin Acikgoz, and the rest of my committee members who guided me throughout my project. I would additionally like to show my appreciation to my cohort, who offered me insight into the technical and theoretical concepts behind the study.

Dedication

I dedicate this paper to my late father, who always taught me to keep feeding my curiosity. Your contagious creativity and commitment to learning live beyond you.

Table of Contents

Abstract	iv
Acknowledgements	v
List of Tables	viii
List of Figures	ix
Introduction	1
Methods	11
Results	23
Discussion	58
References	65
Appendix A: Logistic Regression Results	70
Appendix B: IRB Approval Form	85
Appendix C: Data and Materials Distribution Agreement	86
Vita	90

List of Tables

Table 1. Examples of Item Discrimination Scores	17
Table 2. Variable Operationalization Summary	22
Table 3. Overall Descriptives of Division-Level Variables.....	24
Table 4. Cross-Division-Level Y/N Incident Outcomes.....	28
Table 5. Overall Descriptives of Department-Level Variables	30
Table 6. Department Level Y/N Incident Outcomes	32
Table 7. High/Low Interaction Impact on Incident Probabilities (Discrimination).....	38
Table 8. High/Low Interaction Impact on Incident Probabilities (Safe Behavior Confirmation)	43
Table 9. High/Low Interaction Impact on Incident Probabilities(Free Response Percentage).....	51
Table 10. Summary Table for all Effect Sizes and p-Values of Lags with Maximal Effects.....	57

List of Figures

Figure 1. Simple Effect of Observations on Incidents (Discrimination Score).....	36
Figure 2. Observation/Discrimination Interaction Effects (Odds Ratios)	37
Figure 3. Checklists with Higher Discrimination Scores Predict Lower Incident Probability	39
Figure 4. Simple Effect of Observations on Incidents (Safe-Behavior Confirmation).....	41
Figure 5. Observation/Safe-Behavior Confirmation Interaction Effects	42
Figure 6. Checklists that Confirm Safe Behaviors Predict Lower Incident Probability	44
Figure 7. Simple Effect of Observations on Incidents (Index of Qualitative Variability).....	46
Figure 8. Observation/Index of Qualitative Variability on Interaction Effect on Incidents	47
Figure 9. Simple Effect of Observations on Incidents (Free Response Percentage)	49
Figure 10. Observation/Free Response Interaction Odds Ratios	50
Figure 11. Checklists with Higher Free Response Predict Lower Incident Probability	52
Figure 12. Simple Effect of Observations on Incidents (Free Response Length)	54
Figure 13. Observation/Free Response Length Interaction Effects	55

Introduction

Background

Workplace incidents have a significant financial cost to workers and workplaces alike. The National Safety Council estimates that work injuries cost the United States \$171 billion dollars (National Safety Council, 2020) and take a significant toll on both employees and the companies for which they work. While the financial burden is high, it is arguably even more important to note the amount of physical and mental harm that these accidents bring to workers and their families. In 2019 alone, the United States accrued 2.8 million nonfatal workplace injuries, and fatal workplace accidents claimed lives of five-thousand employees (BLS Census of Fatal Occupational Injuries Summary, 2020; BLS Employer-Reported Workplace Injuries and Illness, 2021).

Behavior Based Safety (BBS), as its name implies, has its roots in the seminal behaviorist publication *The Behavior of Organisms*, where B.F. Skinner lays out a scientific method for observing and modifying behavior (Skinner, 1938). He posits that organisms modify behaviors to get the most optimal outcomes, reducing the performance of behaviors that elicit less pleasant outcomes and increasing the performance of behaviors that elicit more pleasant outcomes. The first notion for applying behavioral principles to workplace safety may have started with Herbert William Heinrich, who emphasized the role of unsafe acts in safety outcomes in his book *Industrial Accident Prevention, a Scientific Approach* in 1931. It was not until the late 70's,

however, that evidence-based research began to lend support for Heinrich's ideas (Sultzer-Azaroff, 1980; Heinrich, 1931).

At its core, BBS is “a process that creates a safety partnership between management and employees that continually focuses people's attention and actions on their own and others' daily safety behavior” (Kabil & Sundararaju, 2019). The company accomplishes this objective through a system of baseline observation assessment, feedback, goal setting, and intervention. At its highest level of an organization, this means incorporating elements of behavioral systems analysis (BSA) and organizational development tools to assess the current state of safe behavior adherence. If there is any discrepancy between actual and desired behaviors, process and systems design can be implemented to reinforce the behaviors necessary for a safe workplace. As employees begin to perform safer workplace behaviors (e.g. double-checking critical process steps, wearing personal protective equipment (PPE) correctly, using proper technique when lifting), downstream safety metrics should begin to improve (ex: fewer accidents due to missing steps, lacking proper PPE, or lifting incorrectly). Once there is an improvement, baseline behavior is assessed again, intervention effectiveness assessed, and then the process starts all over.

To assess and manage behavior, field observations of actual employee behavior must be collected. Most BBS systems accomplish this task through observation checklists: lists that contain items for capturing behavior as it is performed. Behavioral information is typically collected on a peer-to-peer level, although the responsibility for conducting observations may be distributed among all employees, or relegated to a specific employee within a team who is trained to witness and record safe observations (Cooper, 2009). These observations may also be planned, where an observation of a specific task is recorded based on checklist items, or

unplanned, where a worker may only record a behavioral observation after they have witnessed it in the field. Once collected, behavioral information is sent to either directly to a supervisor, to a broader safety information system, or both (Bogard et al., 2015).

Once observations are recorded, a supervisor reviews these checklists and looks for trends in the data, rewarding safe behaviors through positive feedback, and investigating a lack of safe behavior, usually by communicating with the employees involved (McSween, 2003, location 1938-1942). After assessing the context surrounding the lack of safe behavior, the supervisor may clarify proper behavior in the case of a lack of information, or more likely implement some sort of environmental adjustment to promote future safe behaviors. After implementing the intervention, the supervisor continues to monitor trends in observation data to see if the intervention was effective (Geller, 2005).

With proper training, the act of peer-to-peer observation with supervisor review serves four purposes. First, because peers are often closer to the work than anyone else, they are better able to follow the process in question step-by-step. Second, by using the observation checklist, employees are more knowledgeable about what safe behavior entails. Third, ground-level employees are empowered by their ownership of safety, rather than being regulated by top-down policy. Finally, supervisors can highlight employees performing safe behaviors, offering reinforcement that makes future safe behaviors more likely (Bradler et al., 2016).

As BBS is a bottom-up process, there is less policy and overhead monitoring required. This ground-level buy-in also helps foster a values-based culture of safety in the workplace which increases employee participation in evaluating workplace hazards (McSween, 2003, Location 2530-2531). BBS empowers supervisors to focus on positive feedback, making employees more likely to internalize and utilize supervisor suggestions (Kluger & Nir, 2010).

Additionally, as workers become better at recognizing and performing safe behaviors, these behaviors become more automated (DePasquale & Geller, 1999).

BBS also has analytical advantages. Since BBS focuses on desirable, frequent behaviors rather than undesirable, low base-rate injuries, resources can be committed to encouraging specific positive outcomes rather than preventing a plethora of negative ones. Additionally, the organization has greater control over the number of observations conducted (as opposed to the number of incidents that occur), and so the analysis of observations means greater sample sizes and greater statistical power for safety analytics (Dufek et al., 1995). The theories translate to results: BBS management has been shown to be effective at increasing employee participation and reducing incidents in a wide variety of organizations, including on college campuses (Lebbon et al., 2012), manufacturing plants (Sultzer-Azaroff et al., 1990), paper mills (Cooper, 2008), oil refineries (Bogard et al., 2015), construction sites (Choudhry, 2014), and coal mines (Hagge et al., 2017).

BBS is a scientific, iterative process that uses behavioral information collected from the field to create a baseline from which intervention effectiveness is compared. If the baseline and/or intervention measurements are inaccurate, analysis generates inconsistent results which leads to inconsistent program success.

Measurement instrument structure plays a crucial role in collecting accurate information, as poor tools will lead to poor measurement results (Brondolo, 2021). Furthermore, measurement tool structure is a factor contributing to measurement accuracy for which organizations have complete control; while organizations may not be able to control the reporting bias in their workers, they can control the degree to which their tool influences that bias. Advancing the literature on behavioral checklist design will enable organizations to create more accurate

checklists, and therefore superior safety interventions, and provide greater understanding into the factors influencing observation data for BBS researchers.

Observation Checklist Item Quality

The observation checklist is critical to effective BBS intervention, acting as the workhorse which carries the information on which knowledge, reinforcing feedback, and safety analytics are built. As a result, businesses with BBS systems put a great deal of care into creating observation checklists. Subject matter experts (SME's) are consulted to find the most at-risk occupational activities, and then define behaviors which help mitigate these risks, a process known as "pinpointing" (Komaki et al., 1978). Then these pinpoints are compiled into safety observation sheets, which contain pinpointed behaviors specific to a particular project or functional crew (Geller, 2005). The sheet usually contains a list of pinpoints and a fixed-response box for each pinpoint that allows the worker to say "yes/no" or rate the safety of a behavior on a Likert scale. There should also be empty space for observers to fill in contextual narrative data if needed (McSween, 2003, Location 1019-1049).

Despite the importance of pinpointing and observation checklists, and the number of resources in the literature describing best practices for creating these checklists, there has been little empirical work in testing the effectiveness of different text structures of pinpoints (Geller, 2001, 2005; Mayer et al., 2019; Miller, 2006). Specifically, Wirth and Sigurdsson in their 2008 review of safety literature note that there is a lack of research investigating whether specificity in behavioral pinpoints leads to better safety outcomes (Wirth & Sigurdsson, 2008). Their review contrasts Komaki et al.'s suggestion of using pinpoints which capture general classes of behaviors (ex: "body placement") to Geller's recommendation for greater pinpoint specificity (Geller, 2001; Komaki et al., 1978). Komaki's argument for more general pinpoints posits that

the workplace is an exceptionally variable place, and pinpoint flexibility is more important than specificity for identifying behavioral trends. Geller, in contrast, believes that the top few highest-risk behaviors are responsible for most of the occupational accidents in any given work setting, and so those behaviors should be targeted specifically through more detailed checklist pinpoints. While initial analyses indicate that task-specific pinpoints are more effective for incident reduction, the field of pinpoint analysis is still new (Laske, 2020).

Others build off Geller's insights. When making suggestions about pinpoint construction, Ludwig (2018) builds off Johnston and Pennypacker's suggestion to create specific and discrimination pinpoints that describe clear antecedents, behaviors, and consequences during safety-related observations (Ludwig, 2018, pp. 113; Johnston & Pennypacker, 1980). This includes a 'do what, to what, when, and why?' structure. 'Do what' means that the pinpoint contains a verb describing a physical action being taken. 'To what' means that the pinpoint describes a physical object which is receiving the aforementioned action. 'When' means that the pinpoint contains a specific cue that discriminates when the behavior begins and/or ends within a process. 'Why' means that the pinpoint indicates what the behavior is trying to accomplish. By providing all the above information, a pinpoint makes a behavior more discriminant, that is: more specific and differentiable from other behaviors. The recording of discriminant behaviors better-informs workers as to what behavior should be fulfilled, leading to more accurate data collection and more specific reinforcement, and provides better detail for subject matter expert investigation and intervention. Quality checklist items, therefore, should have pinpoints that contain more "do what", "to what", "when", and "why" (discriminant) information in them. An example of a checklist with a highly discriminant pinpoint would be "*Checks (do what) rear-*

view mirror (to what) before reversing the vehicle (when) to ensure the space is clear (why)". A poor example would be "*PPE (what)*".

This study's first hypothesis regarding checklist quality assesses that assumption:

Hypothesis 1a: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklist have more discriminant pinpoints.

Checklist items should also confirm safe behaviors, rather than point out unsafe ones. As stated above, part of BBS's success is that the system can commit resources to specific desirable outcomes, and that workers are more likely to internalize positive feedback. By creating checklist items that confirm safe behaviors, workers are rewarded when the supervisor reviews the checklist and notes good behavior, and organizations are able to track the specifics of what is going right, rather than the ambiguity of everything that can go wrong. A checklist item stating "*employee did not wear proper PPE*" should have less quality than an item stating "*employee wore proper PPE*". The second hypothesis on checklist item quality tests the relationship between safety-confirming questions and safety outcomes:

Hypothesis 1b: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklists contain questions that confirm safe behaviors.

Observation Checklist Response Quality

This study also examines "response quality" and its relationship with safety incidents. Quality checklist responses should be sensitive enough to changes in safe behavior as to be useful for supervisor interventions. There are two reasons that checklists may lack sensitivity: worker motivation or poor checklist item selection.

The workplace provides a variety of stimuli competing for worker resources, including attention and behavior. Along with demands such as observation checklist accuracy, workers must attend to other work tasks and social norms. A worker will choose to perform the tasks that garner the greatest reward, and when other tasks outweigh observation checklist accuracy, this can lead to inaccurate results. While trying to meet a quota, whether explicit or implied, employees may simply mark safe answers and give little context to their observations in order to increase quantity over quality of observations, a phenomenon known as “pencil whipping” (Ludwig, 2014). When employees are trying to impress their supervisor or avoid confrontation with their coworker, they are also more likely to inflate the number of safe observations (Spence & Keeping, 2011).

The result is an influx of observation checklists that all contain the same “safe” answer throughout as employees neglect the environment in order to complete the checklist quickly. With these lists in-hand, supervisors are unable to give quality feedback to their employees and safety professionals are unable to identify the trends in their data critical to a successful safety analytics system. In the current study, response quality is operationalized in two ways based on item type.

First, fixed-response items were evaluated based on variance in responses. Specifically, if an observation checklist with a 1-5 safety scale has many different responses with some 1’s, 2’s, 3’s, 4’s, and 5’s, then responses to that checklist have high variability. If it instead had all 5’s, that checklist’s responses would have no variability. This study assesses how response variability across a checklist acts as a moderator between the number of observations and incident reduction. As variability increases, it indicates that employees are truly evaluating their coworker

during observations, and this should direct peer feedback, and dedicated safety staff to evaluate general trends in safety metrics.

Hypothesis 2: The negative relationship between daily observation checklist counts and daily incident probability will be stronger as fixed-response answer variability within a checklist increases.

Second, as mentioned earlier, textboxes allow observers to further expand on their findings, identifying specific contextual factors that may not be captured in the fixed-response questions (McSween, 2003; Miller, 2006). This additional information allows supervisors to better assess barriers to safe behaviors. Despite the ubiquity of free-text responses in observation checklists, to date there has been no formal assessment of the effectiveness of these responses on incident reduction. This study first examines the mere presence of a written free-text response:

Hypothesis 3a: The negative relationship between observation checklist counts and incident probability will be stronger when the percentage of free-response questions within an observation checklist is greater.

Beyond the existence of narrative free-text, a subject matter expert must be able to understand the text and interpret it in a way that gives useful context to the behavior. Once again, there is a lack of literature specifying what a “quality” narrative response should look like, but research done in other fields (specifically performance appraisals) shows that narrative feedback content analysis is helpful for supplementing quantitative measures of performance (Speer, 2021, 2018). As employees enter more information, the free-text responses should offer greater insight from subject matter experts. The second hypothesis regarding free-text response data tests this proposition:

Hypothesis 3b: The negative relationship between observation checklist counts and incident probability will be stronger when those observation checklists have free response questions with longer responses.

Methods

Data Source

The data for this study comes from a safety database of a large chemical manufacturing company employing over 14,000 employees around the world. The company tracks its occupational safety metrics through a safety reporting information system (SRIS) that gathers information from all of its sites. Despite the worldwide access to this database, different functional parts of the company (divisions) have varying participation rates with the SRIS.

The information submitted by the employees are stored in two different primary datasets. The first, named Audits, contains all information on two main types of leading indicators: inspections and observations. To perform an inspection, dedicated employees conduct a regular, planned exploration of the workplace to find environmental factors that may lead to unsafe conditions. Observations are conducted voluntarily by employees, either on a planned or spontaneous basis, noting down the degree to which a worker follows safe behaviors while performing a task.

The second dataset, named Incidents, contains information on three primary kinds of lagging indicators. The first, hazards, are reported when an employee spots something in the environment that has a potential to cause an injury (example: a dead branch on a tree that may fall). A worker should report a near-miss when they witness an event which almost resulted in an injury (example: a dead branch falls from a tree and almost injures an employee). A true incident

occurs when an event takes place that results in an actual workplace injury (example: a dead branch falls from a tree and hurts an employee).

When an employee enters a report to be saved in either of the above datasets, there are some required pieces of information, including the date of the report, kind of audit/incident, location, and division to which the reporter belongs. Specific to Audits (observations and inspections), the employee must also enter all of the information gathered from the audit, including each individual question and answer.

Two divisions of the company were chosen for this study based on their consistent reporting within the SRIS. Together these divisions employ 1,500 workers. The first division's functional role involves maintenance on machine parts (Maintenance), while the other division focuses on manufacturing (Manufacturing). Workers in each division are regularly introduced to a variety of potential safety hazards, including slips, trips and falls, chemical exposures, environmental releases, hot working conditions, interaction with heavy machinery, and lifting heavy equipment.

A nondisclosure agreement between the host organization and the researchers is included in Appendix A. This agreement outlines the use of nonidentifiable data from the host organization. The host University Internal Review Board (IRB) also approved the use of this data (IRB# 19-0072, Appendix D).

Each division is made up of several departments. This study examined hypotheses on both the departmental and divisional level, depending on the availability of data in each level. Due to statistical power concerns, hypotheses 1 and 3 were investigated using the entire dataset, while Hypothesis 2 was investigated using a single department with the most SRIS information available.

Data Set and Variables

The study examined observation sheets and safety incidents entered into the SRIS by the Maintenance and Manufacturing division employees between 2017 and 2019. For observations, the dataset contains information on a checklist-question level, with each row being a unique question from a unique observation checklist. Each question has data regarding the date it was answered, the division of the employee being observed, checklist number, answer, recommendations, and findings. For incidents, the dataset provides the date that the incident occurred and division and department it occurred in. Weekly employee work hour information was also collected for every employee in each division. This includes total employee hours and overtime hours each week.

Data Transformation and Measures

In order to run our analyses, we first calculated the number of observation forms completed per day in each department. Next, this data was normalized by employee worker hours per week to account for natural variations in the amount of work done or differences in the number of workers. Normalization was performed by first dividing the variable counts by weekly work hours, and then multiplying the result by 200,000. As OSHA has found in their reports, multiplying these counts by 200,000 allows for easier interpretation, as there are often several orders of magnitude more work hours than report counts. Accordingly, observations in this study should be interpreted as “observations per 200,000 weekly work hours”. Normalizations reduce the confounding variable of fluctuating work demands; as more work is being done, there is naturally an increase in both observation counts and incidents. Without normalization, work-hours might act as a confounding variable in the analyses.

Incidents are relatively rare occurrences in the workplace, and often follow a heavily skewed distribution. This means that there are many days with no occurrences, and only a few days per year that have an incident. Skew is quantified in a “skewness” metric, measured on a scale from negative infinity to positive infinity, where the farther away a number gets from zero, the more skewed the distribution is. A high skew is anything above 1.0 or below -1.0. Due to strong skew in the incident/200,000 weekly work hours, this variable was transformed into a dichotomous outcome: 1 represented days that had at least one incident, while 0 represented days without incidents. This new outcome was termed “incident probability” rather than incident count.

Additional variables were added to the dataset to account for the delayed impact of observation checklist counts on incident probability. When an observation checklist is submitted, it takes time for subject matter experts to review the results, have conversations with employees, and implement change. Measuring the impact of observation checklist categories today on incident probability today is therefore unlikely to tell the whole story as it theoretically evaluates a time period where observations checklist counts have minimal impact on incident probability.

In order to find the optimal delay in which observation checklist counts impacts incident occurrence, analyses were conducted across several different time lags. The most impactful relationship would be defined as one with the greatest effect size (defined as the odds that an incident will occur given a one unit increase in observation checklists) and significance (defined as p-value below our alpha, 0.05).

While it is important to measure the relationship between observation checklists on incidents over time, the high variability and strong positive skew of both variables meant that some extreme values may be over-represented in a simple lagged analysis. To account for these

potential extremes, count variables were calculated as rolling sums across their respective lags. Rather than assessing the impact of observations today, yesterday, and the day before, this study measures the impact of observations today, observations across the past two days, and observations across the past three days.

Some of the study's variables (dichotomous, indexed, and percentile) are unsuited for rolling sums, as their values become difficult to interpret when their boundary minimums/maximums are breached (a summed dichotomous variable is useless statistically). Indexed and percentile variables were instead calculated as rolling averages over their given time periods (ex: the one day, two day, and three day averages), while dichotomous variables were calculated as the probability of an even happening across some number of days (ex: the probability of an incident on one day, over two days, or over three days). Finally, predictor variables were also lagged one, two, three... up to seven days. This lag allows the study to examine how observation totals today and the past few days predict incident probability over the course of the next few days.

In sum, the dataset was normalized by every 200,000 weekly work hours. Daily incident totals were turned into a binary daily incident probability variable. To measure the change in incident probability through time, several new variables were created measuring the probability over one, two, three,... up to seven days out. Observations/200,000 weekly work hours were transformed into one day, two day, ..., up to seven day rolling sums. The probability of incidents occurring in the next one to seven days was calculated, and then lagged so that the sum of the previous n number day of normalized observations could be used to calculate the probability of an incident within the next x number of days. Accordingly, this dataset allows for the

measurement of effects from predictors occurring within the last week on outcomes within the next week.

Hypothesis 1a: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklist have more discriminant pinpoints.

Hypothesis 1a was tested using data across both divisions, as there were only 738 unique questions across all departments. To maintain adequate statistical power, all 738 of those questions were assessed across both divisions.

Checklist question quality was assessed across four different factors through hand-coding. This was accomplished by a single coder who assessed whether each checklist item contained behavioral information regarding “do what” meaning that there is a physical action being suggested by the question, “to what” meaning that there is a physical object that is being acted upon in the environment, “when” meaning that there is some temporal or conditional phrase that indicates when this action should be taken within a process, and “for what purpose”, meaning that there is some desired outcome to be achieved. No measure of inter-rater reliability was assessed, as items were only measured by a single rater. A good example of a behavioral observation may be “*seals off (do what) energy source (to what) after checking with supervisor (when) to ensure no accidental release (why)*”. A particularly poor example may be “*check energy*”. For a list of examples, please refer to Table 1.

Table 1*Examples of Item Discrimination Scores*

Item	Do What?	To What?	When?	Why?
High Actionability Items				
How did you check the vehicle for stability before raising to working height?	Check	The vehicle	Before raising to working height	For stability
When repairing steam leaks, how did you position body out of line of fire?	Position	Body	When repairing steam leaks	Out of line of fire
Low Actionability Items				
Rushed	Rushed	N/A	N/A	N/A
Using Hazard Lights	Using	Hazard Lights	N/A	N/A

Information on the degree that a pinpoint was discriminant was operationalized with a “Discrimination Score.” To generate this score, a list of all questions was generated. Each question was rated as behavior- or non-behavior based on whether the item was an action performed by a worker or a condition of the environment (i.e., hazard). A behavioral question got an additional point of Discrimination Score for each of the four indicators of discriminant behavior (Do What? To What? When? Why?), with a minimum of 0 and a maximum of 4. Average Discrimination Score for a given checklist was calculated by dividing the total Discrimination score across all questions by the total number of questions for that checklist. Then the average Discrimination Score for a checklist across an entire day was calculated by adding all of the average checklist Discrimination Score’s together for a day and dividing by the number of observation checklists for that day. To account for employee hours, that variable was divided by the weekly employee hours and multiplied by 200,000. The final variable was a normalized measure of average Discrimination Score per observation checklist per day.

In summary, the operationalization of items with discriminant pinpoints was the sum of the normalized average number of questions per observation checklist per day over the course of one through seven days.

Hypothesis 1b: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklists contain questions that confirm safe behaviors.

A single coder evaluated all checklist items. That rater first marked each item as describing a behavior or not, then marked each item as confirming safe behaviors or not. An example of a safe-behavior confirming item would be “*Wore PPE Correctly*”. To calculate the average number of safe-behavior confirming items per observation sheet per day, the number of

safe-behavior confirming items within an observation sheet was added together and then divided by the total number of items within that check sheet. The average number of safe-behavior confirming items was then added together and divided by the total number of observations for a given day. This variable was then normalized by dividing by the total number of weekly hours and multiplying by 200,000. Finally, this variable was aggregated across time as a rolling sum.

In summary, the operationalization of safe-behavior affirming items was the sum of the normalized average number of questions per observation checklist per day over the course of one through seven days.

Hypothesis 2: The negative relationship between daily observation checklist counts and daily incident probability will be stronger as fixed-response answer variability within a checklist increases.

The Index of Qualitative Variation (IQV), as the name implies, is a metric for measuring the variation within qualitative variables. The value ranges from 0 to 1, with greater numbers meaning more variation. IQV is calculated with the following equation.

$$IQV = K(100^2 - \sum Pct^2) / 100^2(K - 1)$$

K is the number of categories in the distribution and Pct is the percentage representation of each category in the distribution. If only one category is represented, then one percentage becomes 100 and all other percentages become 0. Squaring that value puts 10000 in the numerator, and IQV becomes 0. Oppositely, if all variables are represented equally, the numerator becomes equal to the denominator, and IQV becomes 1.

To determine IQV within this dataset, first a coder evaluated each fixed item and determined the number and kinds of possible answers to that item, which would be representative of K. As an example, there may be several questions within a checklist that have the response “Yes, No, NA”

as available response options. Another group of questions may have “Safe, Unsafe, NA” as available response options. These questions would have the same number of possible responses, but different response options. Then questions with the same number and types of fixed response options were aggregated within a checklist such that the percentage of each possible answer was calculated. Those percentages were exponentiated, then summed, subtracted from 10,000, multiplied by K, and then divided by 10,000 times K – 1. The result was an average IQV per free-item response group (kinds of responses and number of responses). IQV across all groups was averaged to arrive at an average IQV across one checklist. That value was multiplied by 100 to allow for easier interpretation of odds ratios from logistic regression models. While IQV typically ranges from 0 to 1, these values range from 0 to 100. Daily IQV average was determined by taking the average IQV across all observation checklists across a day. IQV was aggregated across time by taking the average IQV across each time period.

In summary, the final operationalization of fixed-response variability was an n-day average of IQV per observation checklist per day.

Hypothesis 3a: The negative relationship between observation checklist counts and incident probability will be stronger when the percentage of free-response questions within an observation checklist is greater.

To preserve statistical power, Hypothesis 3 was analyzed across divisions, as there were only 65 different templates across the dataset. There were only 738 unique questions across all departments, and of those 125 were free response questions that included behavior-irrelevant inquiries into the location of the observation checklist and the task being performed. Of those, only 70 questions were behavior-based in nature. A single coder rated each question as either free- or fixed- response. The number of free response vs total response questions were counted

across each template version, and a percentage free-response questions per checklist was calculated. That checklist average was average across the daily level. To assess average free response across several periods, a rolling average of the past one to seven days was calculated by averaging the metric and then lagging its values. This value was multiplied by 100 to make odds ratio interpretations easier.

In summary, the presence of free-response questions was operationalized as the rolling average percent of free response questions per observation checklist per day across one through seven days.

Hypothesis 3b: The negative relationship between observation checklist counts and incident probability will be stronger when those observation checklists have free response questions with longer responses.

Free response answer length was calculated by counting the characters in each free response question response, averaging them across each observation checklist, and then getting the daily average of free response answer length. Then past one to seven day answer lengths were calculated by computing the rolling averages across one to seven days.

In summary, free-response length was operationalized as the average free-response answer length per observation checklist per day, averaged across one to seven days.

Operationalization Summary

As this study contains a broad array of operationalizations, Table 2 provides a summary of all operationalizations. The conceptual variable is listed in the “variable” column. Operational units display the transformations performed on to obtain the quantitative variable needed for analysis. The aggregation-over-time method shows how each variable was aggregated across one to seven days.

Table 2*Variable Operationalization
Summary*

Hypothesis	Variable	Operational Unit	Aggregation-Over-Time Method
1A	Discriminant Pinpoints	Average Discrimination Score per Observation Checklist Per Day	Sum
1B	Safe Behavior Confirmation	Average Number of Items Confirming Safe Behaviors per Observation Checklist Per Day	Sum
2	Fixed-Item Response Variation	Average Index of Variation per Observation Checklist per Day, multiplied by 100	Average
3A	Free Response Questions	Average percentage of free response questions per observation checklist per day, multiplied by 100	Average
3B	Free Response Length	Average free response length per observation checklist per day	Average
Overall Tests	Incidents	Occurrence of at least on incident each day	Occurrence

Results

Descriptive Statistics

Cross-Division Statistics

Table 3 shows the cross-division descriptive statistics for the one- and seven-day aggregations of the normalized continuous variables. Missing values come from performing rolling-sum and lag transformations. Incidents were skewed (Skewness > 1) across all rolling sums, which is the primary reason this study conducted logistic regression analyses, as logistic regression helps account for skewness in the outcome variable by changing it into a binary variable. Individual measures for the first hypothesis (H1a; “Discrimination Score”) show skew, but after aggregation the total Discrimination Score is not skewed. The sizable maximums in occurrences comes from the normalization equation mentioned above, which was calculated by dividing by weekly hours and multiplying by 200,000, a standard set by OSHA containing a year’s worth of worker data per 100 employees (100 employees working 40 hours per week, 50 weeks per year).

Table 4 in shows descriptives for the binary outcome variables. Only 13.5% of days contained at least one incident, and 63.1% of seven-day time periods contained at least one incident.

Table 3*Overall Descriptives of Division-Level Variables*

Variable (Normalized)	N	Missing	Mean	Median	SD	Coefficient of Variation
Summed Variables						
Incident 1-Day Total	1094	0	5.00	0.00	14.80	2.96
Incident 7-Day Total	1094	0	34.98	22.14	41.08	1.17
Observation 1-Day Total	1093	1	86.94	82.83	49.69	0.57
Observation 7-Day Total	1081	13	610.62	590.44	161.29	0.26
Do What 1-Day Total	1093	1	406.07	372.65	277.67	0.68
Do What 7-Day Total	1081	13	2854.62	2714.47	882.22	0.31
To What 1-Day Total	1093	1	389.79	362.96	262.05	0.67
To What 7-Day Total	1081	13	2739.61	2614.62	819.43	0.30
When 1-Day Total	1093	1	60.41	43.48	61.09	1.01
When 7-Day Total	1081	13	425.01	328.46	241.85	0.57
Why 1-Day Total	1093	1	15.82	0.00	24.21	1.53
Why 7-Day Total	1081	13	110.98	116.70	109.51	0.99
Affirmation 1-Day Total	1093	1	0.31	0.00	0.39	1.29
Affirmation 7-Day Total	1081	13	2.14	1.75	1.36	0.64
Weekly Work Hours	1094	0	54605.92	56008.13	7384.20	0.14

Table 3 (continued)

Overall Descriptives of Division-Level Variables

Variable (Normalized)	N	Missing	Mean	Median	SD	Coefficient of Variation
		Averaged Variables				
Actionability Score 1-Day Total	1093	1	18.37	0.00	27.11	1.48
Actionability Score 7-Day Total	1081	13	18.32	1.78	22.37	1.22
Percent Free-Response 1-Day Average	1093	1	17.74	0.00	26.26	1.48
Percent Free-Response 7-Day Average	1081	13	17.69	1.77	21.57	1.22
Free Response Question Answer Length 7-Day Average	1093	1	2.37	0.00	2.65	1.12
Free Response Question Answer Length 7-Day Average	1081	13	2.39	1.97	1.89	0.79

Table 3 (continued)*Overall Descriptives of Division-Level Variables*

Min	Max	Skewness		Percentiles		
		Skewness	SE	25th	50th	75th
Summed Variables						
0.00	104.69	3.60	0.07	0.00	0.00	0.00
0.00	257.64	1.59	0.07	0.00	22.14	52.68
0.00	253.50	0.67	0.07	49.71	82.83	116.63
0.00	1157.47	0.46	0.07	488.85	590.44	713.47
0.00	1444.34	0.79	0.07	173.88	372.65	563.79
0.00	5894.22	0.72	0.07	2242.90	2714.47	3312.05
0.00	1400.20	0.75	0.07	170.43	362.96	542.47
0.00	5566.33	0.68	0.07	2176.86	2614.62	3176.63
0.00	381.47	1.32	0.07	10.39	43.48	92.37
0.00	1256.53	0.86	0.07	236.16	328.46	614.62
0.00	117.80	1.58	0.07	0.00	0.00	27.57
0.00	439.04	0.52	0.07	0.00	116.70	190.03
0.00	1.91	1.21	0.07	0.00	0.00	0.51
0.00	6.89	0.78	0.07	1.17	1.75	3.01
22101.00	70163.88	-1.60	0.07	52972.33	56008.13	57914.90

Table 3 (continued)

Overall Descriptives of Division-Level Variables

Min	Max	Skewness		Percentiles		
		Skewness	SE	25th	50th	75th
Averaged Variables						
0.00	99.01	0.72	0.07	26.23	34.41	41.47
0.00	435.93	0.45	0.07	217.08	238.75	262.18
0.00	100.00	1.23	0.07	0.00	0.00	37.04
0.00	91.43	0.78	0.07	0.45	1.78	40.02
0.00	14.88	0.57	0.07	0.00	0.00	4.80
0.00	6.45	0.19	0.07	0.66	1.97	4.20

Table 4

*Cross-Division-Level Y/N
Incident Outcomes*

Variable	Missing	Mean	Sum	Coefficient of Variation	SD
1-Day Incident Occurrences	0	0.135	148	2.533	0.342
7-Day Incident Occurences	0	0.631	690	0.765	0.483

The coefficient of variation is a quick way to compare overall variation in the variables (equation is $C = SD/Mean$). As datasets are aggregated to larger numbers, the coefficient of variation decreases, meaning that variation is also decreasing.

Department-Specific Statistics

Table 5 displays the continuous data for the individual department used to answer Hypothesis 2. Data is missing from the dataset for the same reason stated above. Skew in the outcome variable across all rolling sums is even more apparent in this dataset (15.061 for 1-day sums and 5.466 for 7-day sums). While observations and 7-day IQV averages were slightly skewed, this difference only slightly above 1 and below -1 (1.388 and -1.246 respectively). Aggregation has a similar effect on this dataset, where overall variation decreases.

Table 6 displays the binary data for the individual department. Incident occurrence is, predictably, even more rare, with only 0.6% of days containing any incidents, and 3.9% of 7-day time periods containing at least one incident.

Table 5*Overall Descriptive Statistics for Department-Level Variables*

Variable (Normalized)	N	Missing	Mean	Median	SD	Coefficient of Variation
<i>Summed Variables</i>						
Incident 1-Day Total	1075	0.00	0.11	0.00	1.50	14.02
Incident 7-Day Total	1075	0.00	0.75	0.00	3.92	5.21
Observation 1-Day Total	1075	0.00	141.88	117.85	115.33	0.81
Observation 7-Day Total	1075	0.00	991.34	935.44	452.90	0.46
Weekly Hours	1075	0.00	10154.77	9121.92	3103.28	0.31
<i>Averaged Variables</i>						
Index of Qualitative Variation 1-Day Average	1075	0.00	4.91	5.38	2.06	0.42
Index of Qualitative Variation 7-Day Average	1075	0.00	4.89	5.11	1.07	0.22

Table 5 (continued)

Overall Descriptive Statistics for Department-Level Variables

Minimum	Maximum	Skewness		Percentiles		
		Skewness	SE	25th	50th	75th
Summed Variables						
0.00	28.24	15.06	0.07	0.00	0.00	0.00
0.00	28.24	5.47	0.07	0.00	0.00	0.00
0.00	756.25	1.39	0.07	55.48	117.85	198.92
0.00	2639.36	0.72	0.07	662.91	935.44	1235.44
3333.00	16374.11	0.62	0.07	8162.31	9121.92	11882.07
0.00	8.17	-1.04	0.07	3.82	5.38	6.46
0.00	6.80	-1.25	0.07	4.37	5.11	5.61

Table 6*Department-Level Y/N Incident Outcomes*

Variable (Normalized)	N	Missing	Mean	Sum	Coefficient of Variation	SD	Minimum	Maximum
1-Day Incident Occurrences	1075	0	0.00558	6	13.35	0.0745	0	1
7-Day Incident Occurrences	1075	0	0.03907	42	4.96	0.1939	0	1

Hypothesis 1a: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklist have more discriminant pinpoints.

Average daily question quality and observation counts across the past seven days, as well as their interaction, were entered across forty-nine binary logistic regression models predicting the probability of an incident across the next one to seven days. These analyses were conducted on the cross-division level. The lag with maximal effect (LME) for observations impacting incident probability was observation counts for the past two days predicting the probability of an incident over the course of the next week (OR = 1.006, $p = .035$). The corresponding interaction effect was also significant (OR < 1.000, $p = .023$), and the overall model with both predictors and their interaction was significant for reducing the error variance in the outcome ($R^2_N = 0.0098$, $\chi^2 = 7.82$, $p = 0.050$).

As shown in Table 7 and Figure 1-2, when the average observation item quality across the past two days is low, low numbers of observation checklists means that there is only a slight probability (0.56) of an incident over the next seven days, but this probability increases as numbers of low-quality observation checklists increases (up to 0.74). The reverse is true when observation item quality is high (0.73 to 0.54). These results support Hypothesis 1a.

The effect size measures for these results are presented as odds ratios. An odds ratio is the estimated increase in the log-odds of the *outcome per one unit increase of the predictor* (Szumilas, 2010). Odds and probability are two ways to measure the chance of a future binary outcome's occurrence. Probability compares the number of events for which an outcome occurs to the total number of possible events. As an example, the probability of rolling a 1 on a die with six sides is 1 (the number of possible events for which the outcome in question occurs) out of 6

(the number of events possible), or 1/6. The odds of an outcome occurring instead compare the number of events for which an outcome occurs to the number of events to which an outcome does not occur. For the same die example, the odds would be 1 (the number of possible events for which the outcome in question occurs) out of 5 (the number of possible events for which the outcome does not occur).

An odds ratio compares two odds to one another, and an odds ratio in a logistic regression examines how a one point increase of the predictor impacts the likelihood of the outcome. As an example, if a predictor has an odds ratio of 2.0 when predicting an outcome, that means that a one point increase in the predictor will make the outcome twice as likely. As a result, an odds ratio of 1.0 means that there is no effect, as a likelihood multiplied by 1.0 remains the same.

Figure 1 is a Logistic-Lag graph, displaying all of the Odds Ratio outputs for all combinations of 1-7 day aggregations of predictors and 1-7 day aggregations of outcomes. On the x-axis is a measure of Odds Ratios (OR). On the left y-axis, Figure 1 displays the 1- through 7- day aggregations of incident occurrence. The right y-axis displays the 1- through 7-day aggregations of the predictors. Significant relationships are highlighted in black. In Figure 1, the black line is in the 2nd box as indicated by the right y-axis, meaning that it corresponds to the 2-day rolling total of observations and the 2-day rolling average of Discrimination Score. It rests on the 7th line as indicated by the left y-axis, meaning that it corresponds to the 7-day rolling aggregation of incident occurrence. In total, this means that the 2-day rolling total of observations predicting outcomes, while holding both 2-day average Discrimination Score and the interaction constant, were significant in their contribution to predicting incidents within the next 7 days, and made incidents more likely to occur.

Figure 3 is a marginal means plot. On the x-axis of this plot is the predictor variable, in this case the 2-day rolling sums of normalized observations. This axis contains all values from the minimum to maximum number of that predictor variable. On the y-axis is the probability of the outcome variable occurring, in this case the probability of an incident within the next week. The plot contains three lines, one for high (one standard deviation above the mean), medium (the mean) and low (one standard deviation below the mean) values of the interaction term, in this case, the interaction term is Discrimination Score. This graph displays all predicted probabilities of an incident in the next seven days for all values of low, medium, and highly discriminant observation checklists. As highly discriminant ($SD + 1$) observation checklists increase, incident probability decreases. As low-discriminant ($SD - 1$) observation checklists increase, incident probability increases. Table 7 displays specific probability values for high, medium, and low discrimination observation checklists for the 0th, 25th, 50th, 75th, and 100th percentile of observation counts.

Figure 1

*Simple Effect of Observations on Incidents
(Discrimination Score)*

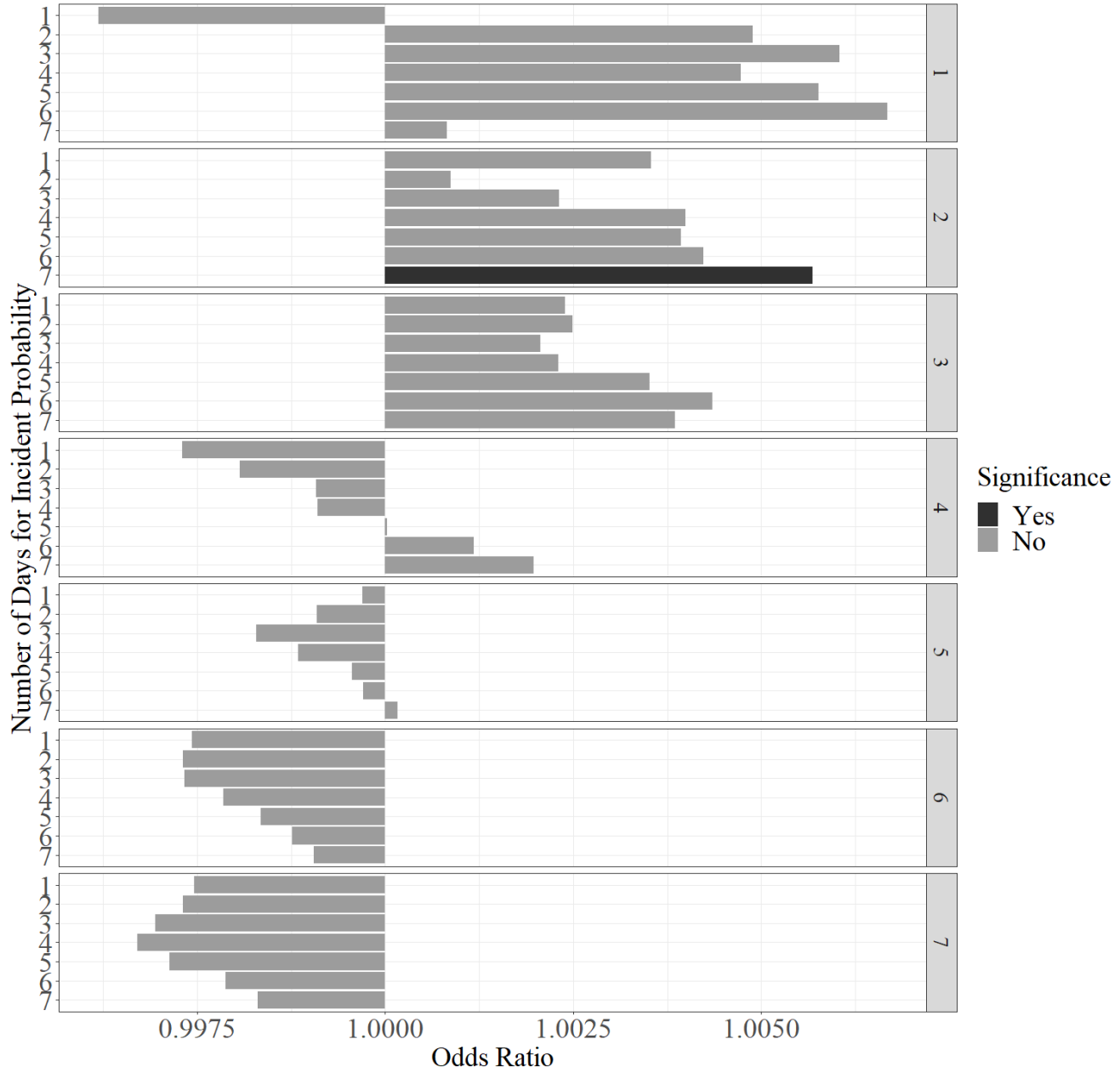


Figure 2

Observation/Discrimination Interaction Effects

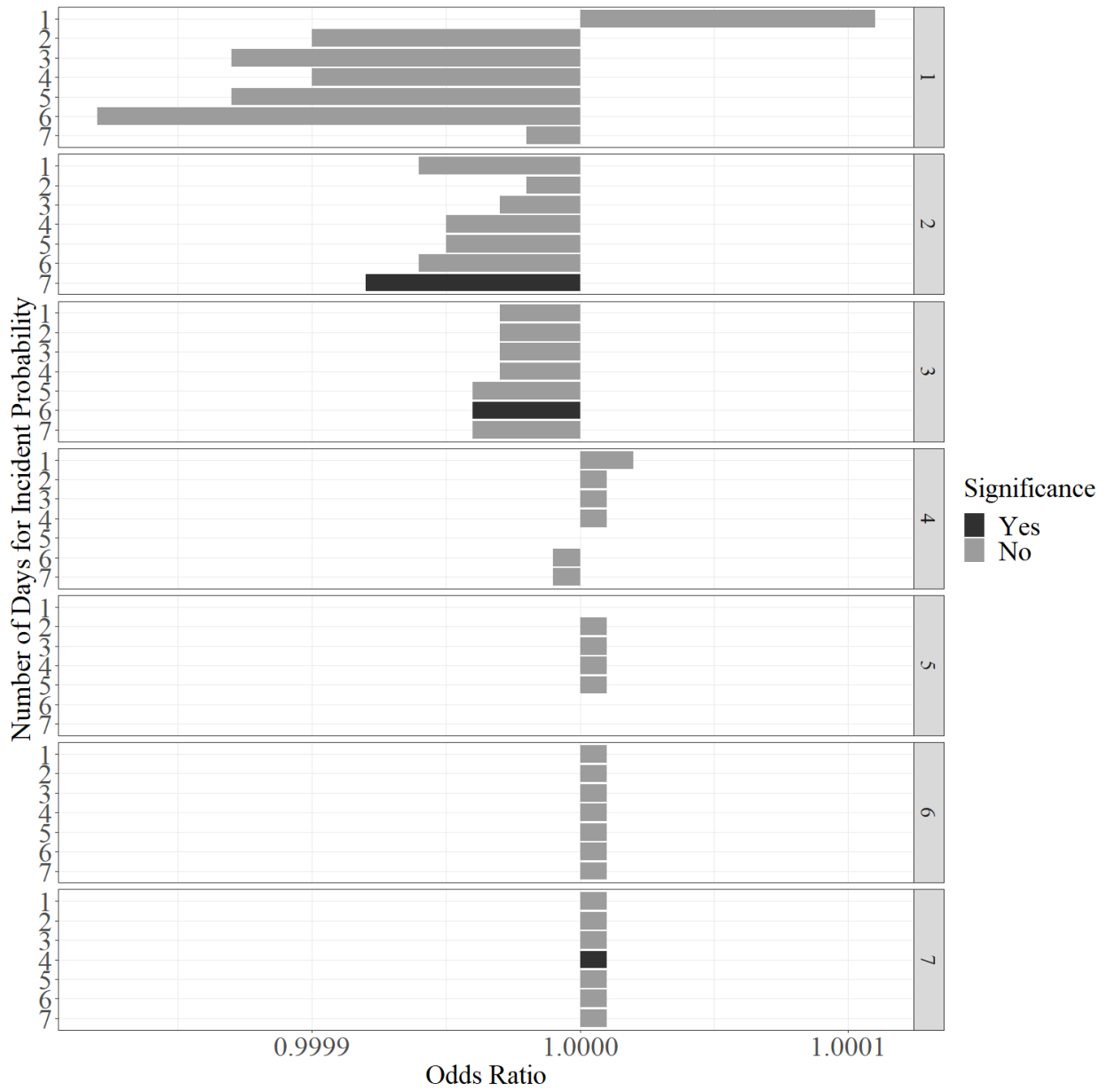


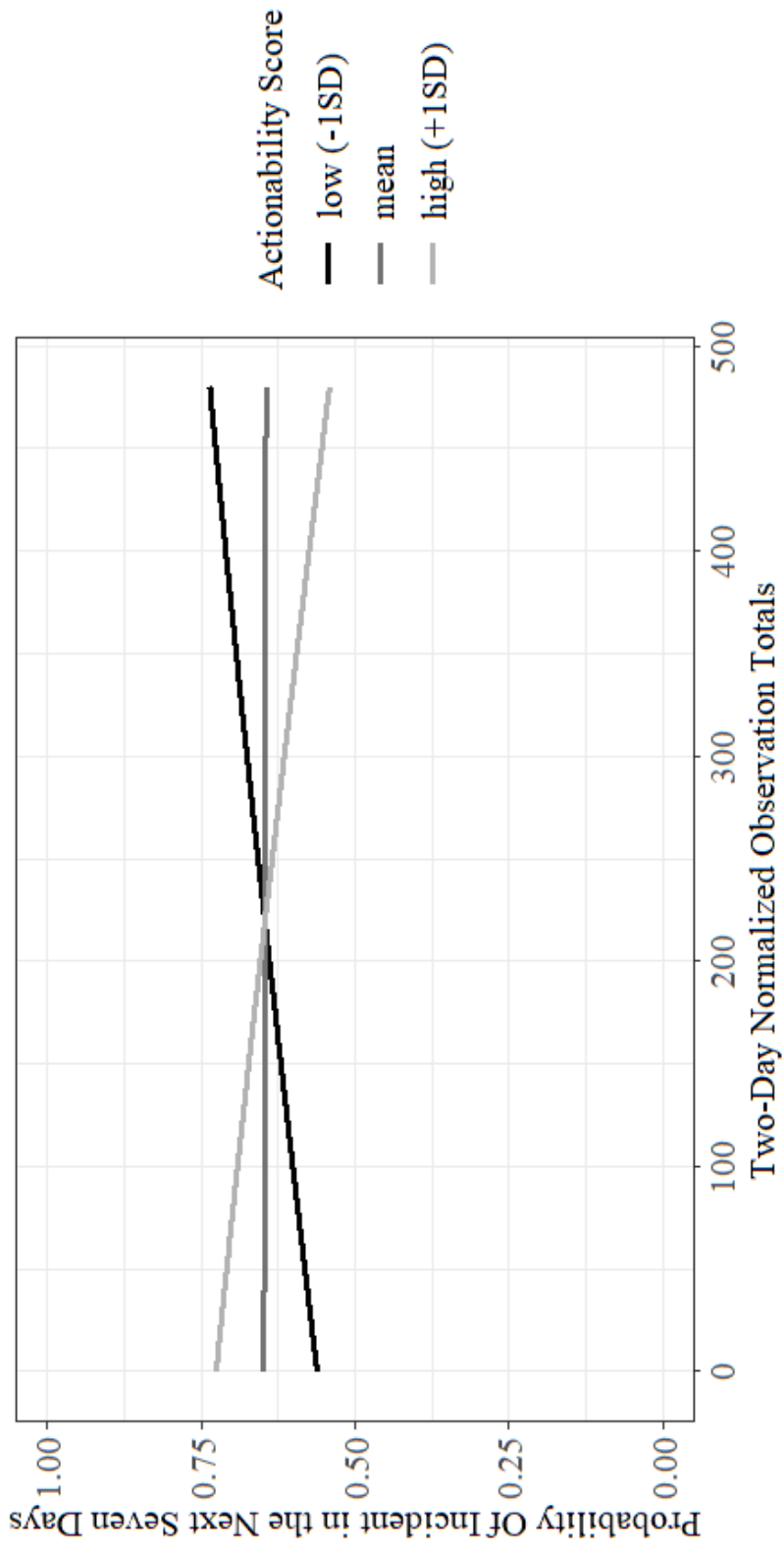
Table 7

*High/Low Interaction Impact on Incident Probabilities
(Discrimination)*

Count	Percentile				
	0th	25th	50th	75th	100th
	0	120	240	360	480
Actionability Value					
SD - 1	0.56	0.61	0.65	0.70	0.74
SD + 1	0.73	0.68	0.64	0.59	0.54

Figure 3

Checklists with Higher Discrimination Scores Predict Lower Incident Probability



Hypothesis 1b: The negative relationship between daily normalized observation checklist counts and daily incident probability will be stronger when observation checklists contain questions that confirm safe behaviors.

The interaction between number of questions confirming safe behavior per day and observation checklist count per day was evaluated across forty-nine different regression models. Observation checklist counts had the strongest impact on incident probability when comparing observations over the past two days predicting incident probability over the next two days (OR = 1.002, $p = .049$). The overall model fit for this test was significant ($R^2_N = 0.0205$, $\chi^2 = 15.2$, $p = 0.002$), and the interaction effect for this LME was also significant (OR = .996, $p = .004$). Table 8 and Figures 4-5 display that when observation counts over the past two days contain fewer safety-confirming questions, there is a greater probability of an incident occurring over the next two days (0.15 to 0.53). When observation counts over the past two days contain more safety-confirming questions, there is a lower probability of an incident occurring over the next few days (0.38 to 0.12). See Table 8 and Figure 6 for more information.

Figure 4

*Simple Effect of Observations on Incidents
(Safe-Behavior Confirmation)*

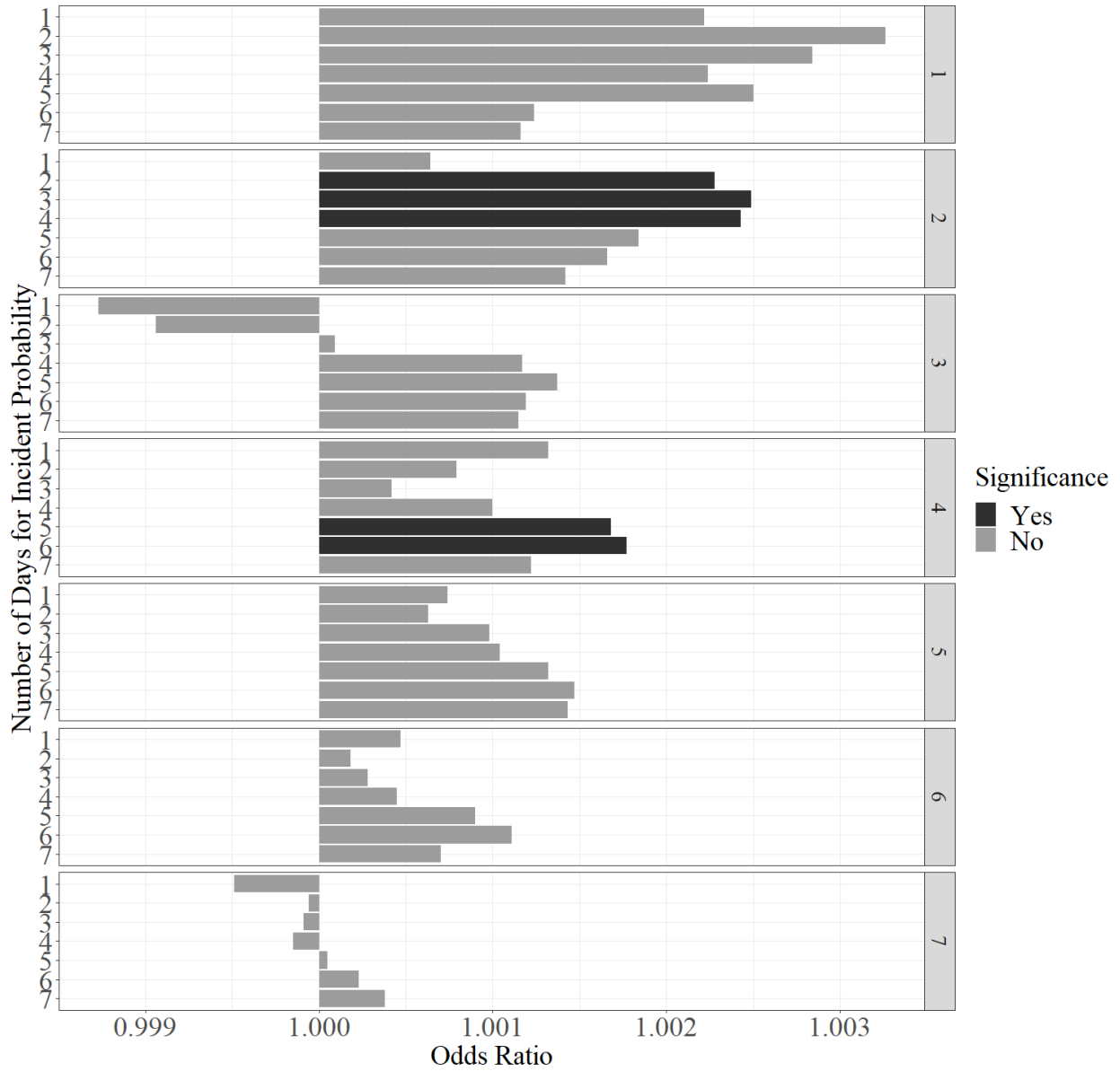


Figure 5

Observation/Safe-Behavior Confirmation Interaction Effects

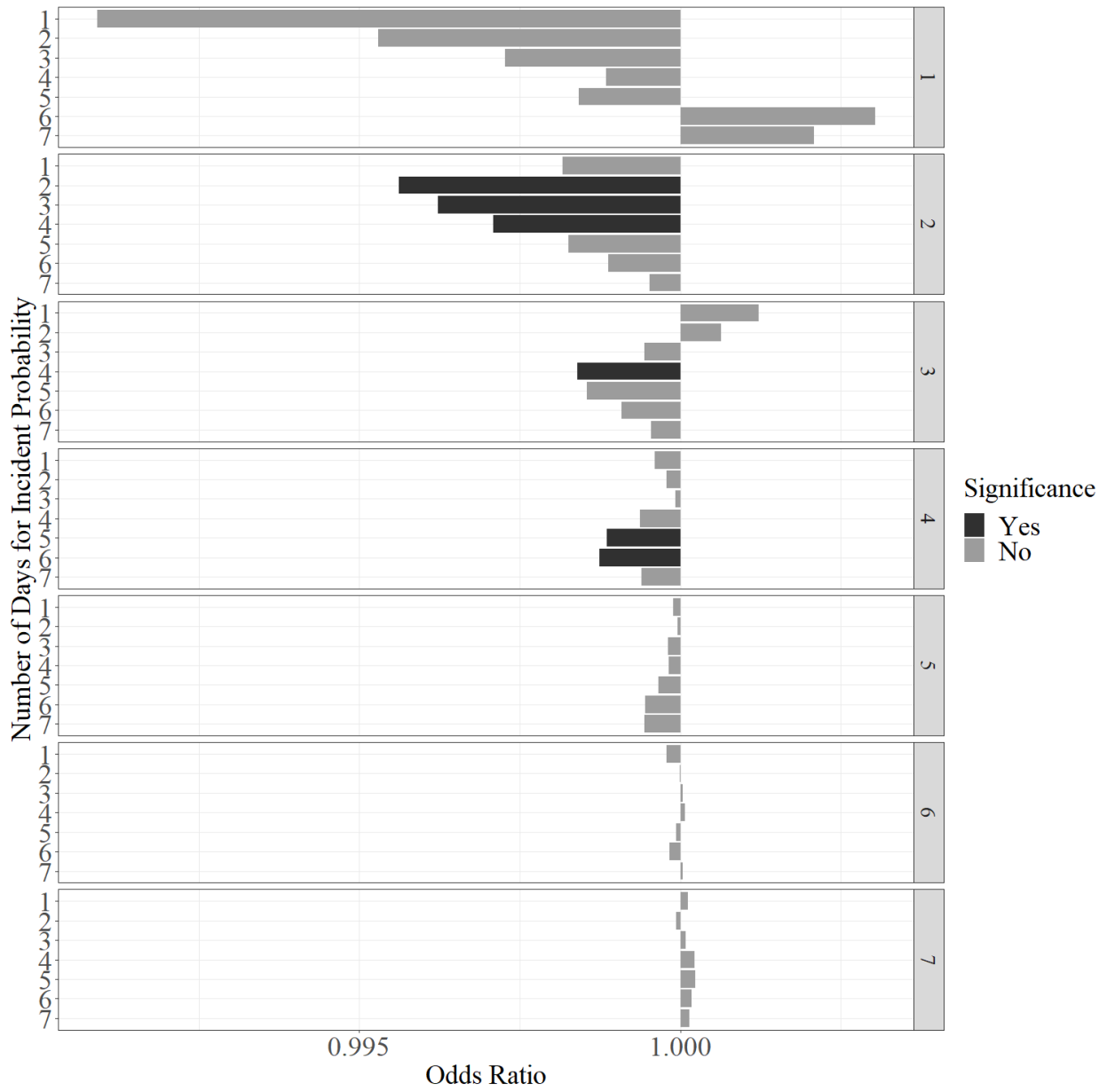


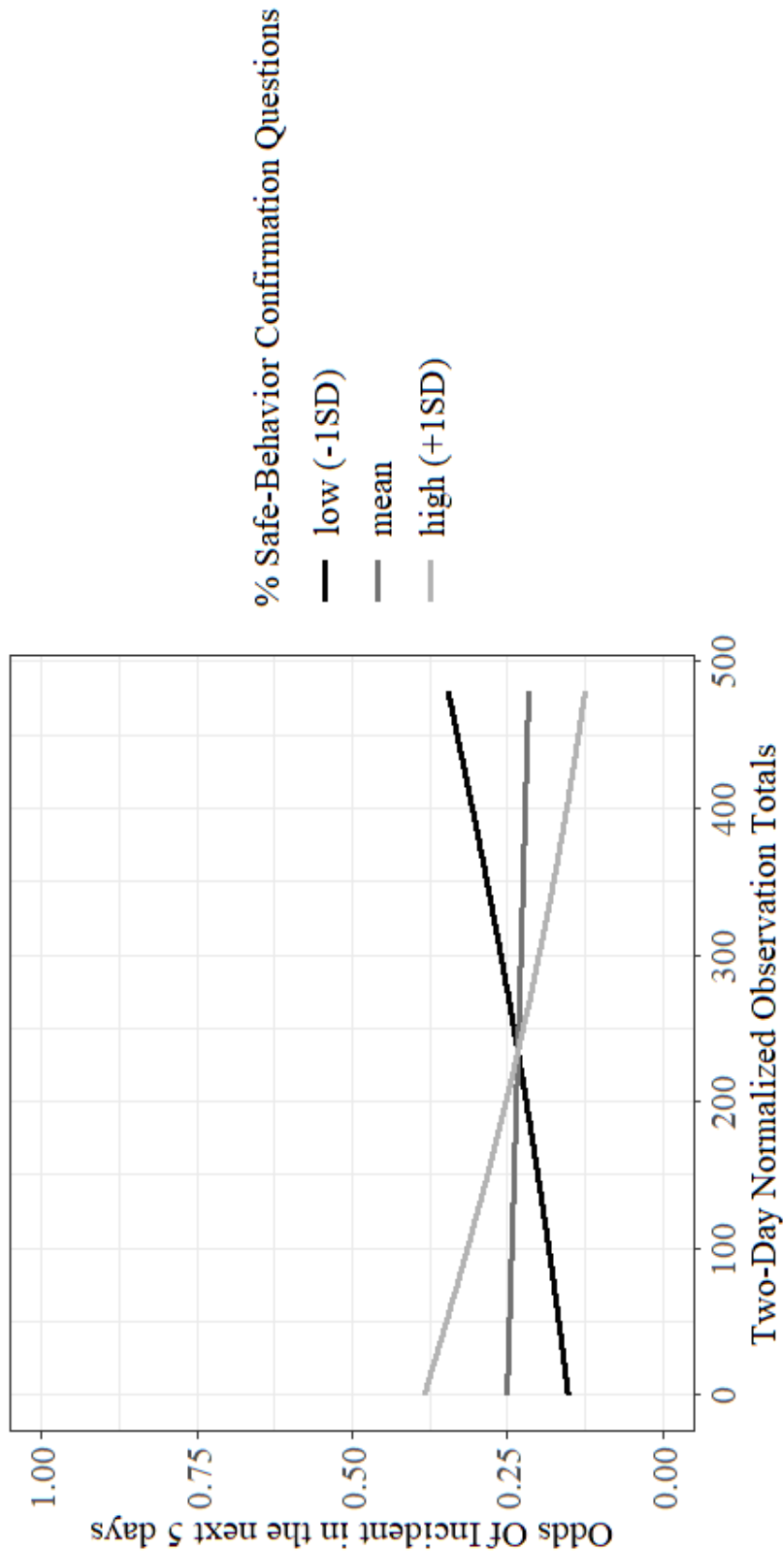
Table 8

*High/Low Interaction Impact on Incident Probabilities
(Safe Behavior Confirmation)*

Count	0th	25th	50th	75th	100th
	0	120	240	360	480
Confirmation Value					
SD - 1	0.15	0.19	0.23	0.29	0.53
SD + 1	0.38	0.30	0.23	0.17	0.12

Figure 6

Checklists that Confirm Safe Behaviors Predict Lower Incident Probability



Hypothesis 2: The negative relationship between daily observation checklist counts and daily incident probability will be stronger as fixed-response answer variability within a checklist increases.

To find the maximum delayed relationship between observation rolling sums and incident occurrence, forty-nine logistic regressions were run with observation checklist rolling sums across each of the past seven days and average IQV across each of the past seven days as predictors and the probability of incident occurrence across each of the next seven days as outcomes. The outcome-predictor lag pair with the greatest significance and effect between observation counts and incident probability was selected as the baseline relationship before assessing the interaction impact. The LME was identified by first eliminating any regressions that did not result in a significant relationship between observation counts and incident probability. From among the significant results, the maximum odds ratio was selected, and then interaction effects were checked for significance/effect.

As shown in Figures 7-8, The four-day rolling total of observations within the department had a LME on the next-day probability of an incident in that same department (OR = 1.011, p = .032). After evaluating the impact of IQV, observation counts, and the interaction between them, the overall chi-squared model failed to significantly differ from a null model with no predictors ($R^2_N = 0.0559$, $\chi^2 = 4.01$, p = 0.260). This means that, regardless of independent variable contributions to the model, the model was not helpful for predicting incident probability beyond a null model.

Figure 7

Simple Effect of Observations on Incidents (Index of Qualitative Variability)

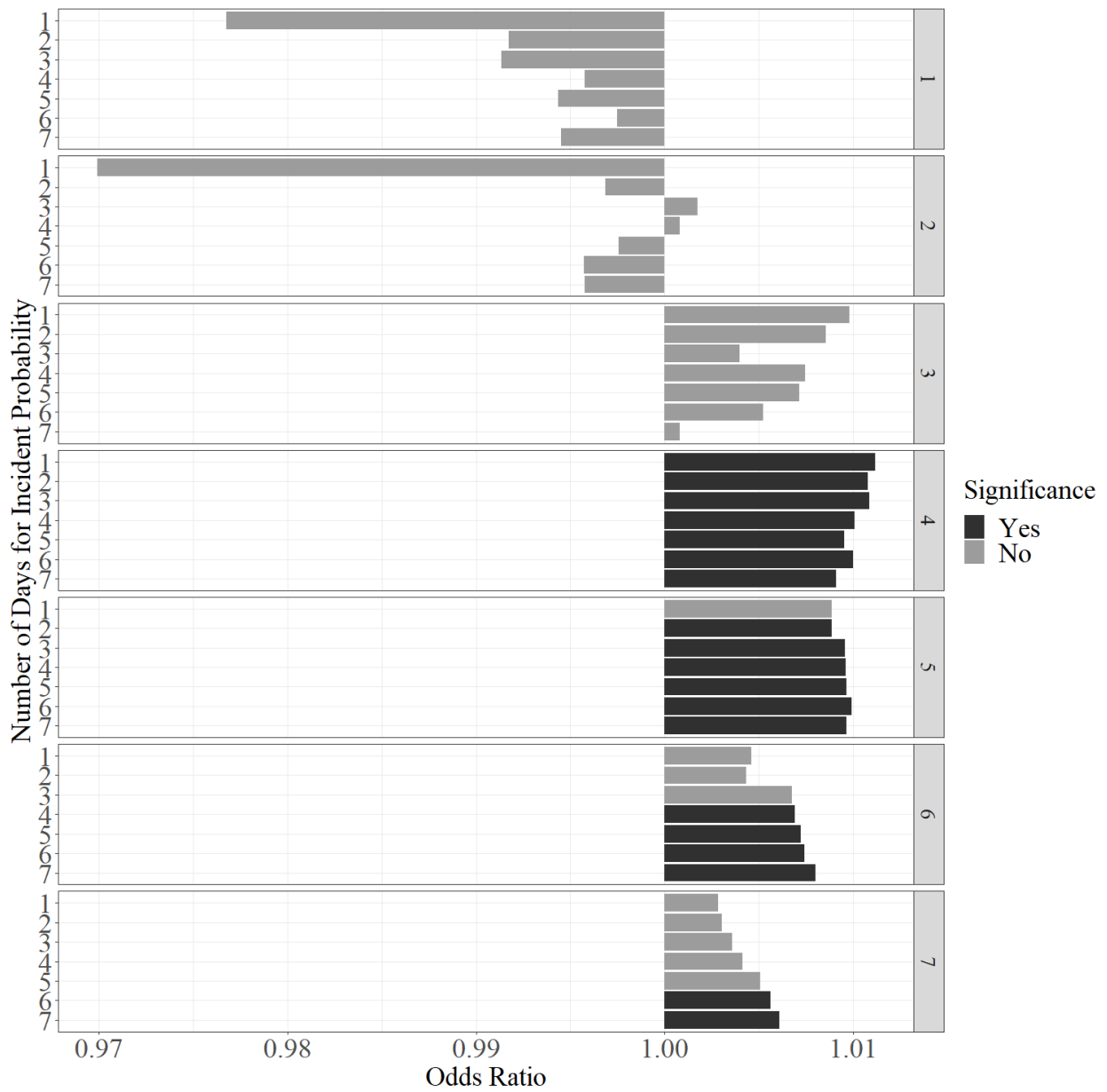
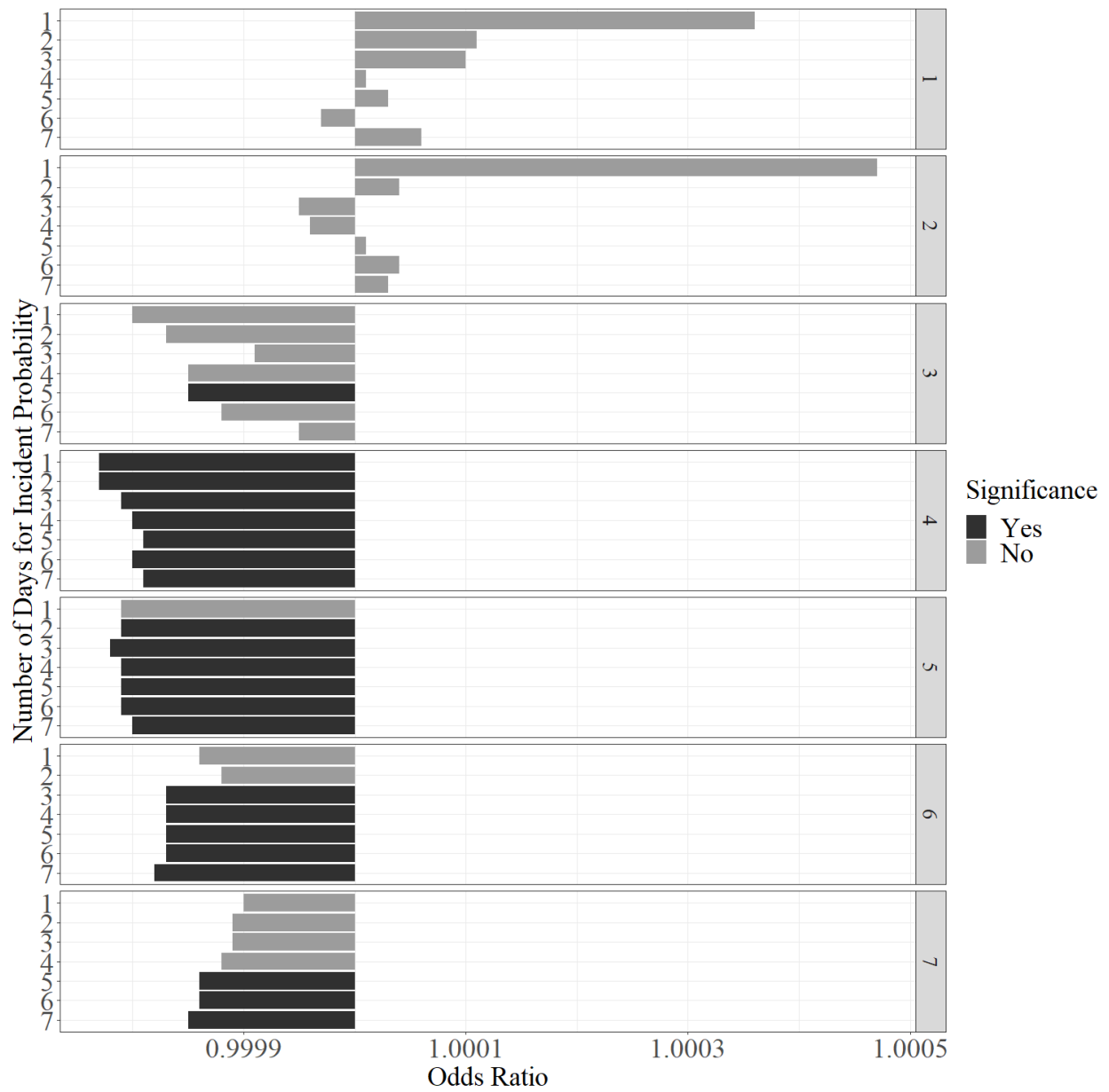


Figure 8

Observation/Index of Qualitative Variability on Interaction Effect on Incidents



Hypothesis 3a: The negative relationship between observation checklist counts and incident probability will be stronger when the percentage of free-response questions within an observation checklist is greater.

Forty-nine logistic regression models were built with each model containing the sum of each of the past seven days of observation counts predicting the probability of an incident in each of the next seven days, with the past seven-day average percentage of free response answers as a moderator.

Observation counts today were best at predicting incident occurrence across the next five days (OR = 1.004, $p = .006$). The overall model test ($R^2_N = 0.0352$, $\chi^2 = 29.3$, $p = < 0.001$) and interaction effect for this LME was also significant (OR < 1.000, $p = .021$), meaning that when the free response percentage is low, an increase in observation checklists today leads to an increase in the probability of an incident in the next five days (0.470 to 0.769). When the free response percentage is high, incident probability is instead reduced (0.470 to 0.387). See Table 9 and Table 9 for more details.

Figure 9

Simple Effect of Observations on Incidents (Free Response Percentage)

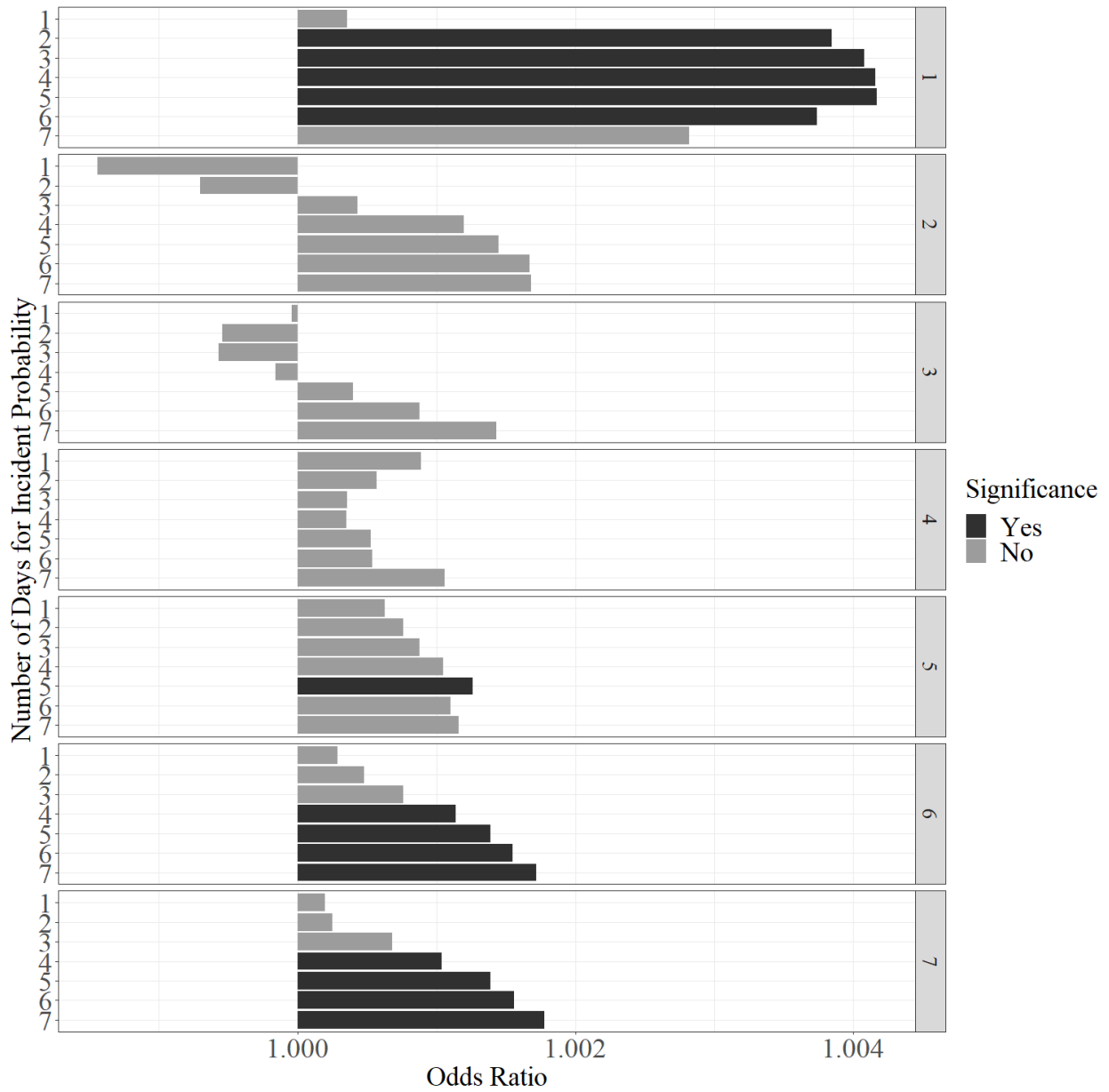


Figure 10

Observation/Free Response Interaction Odds Ratios

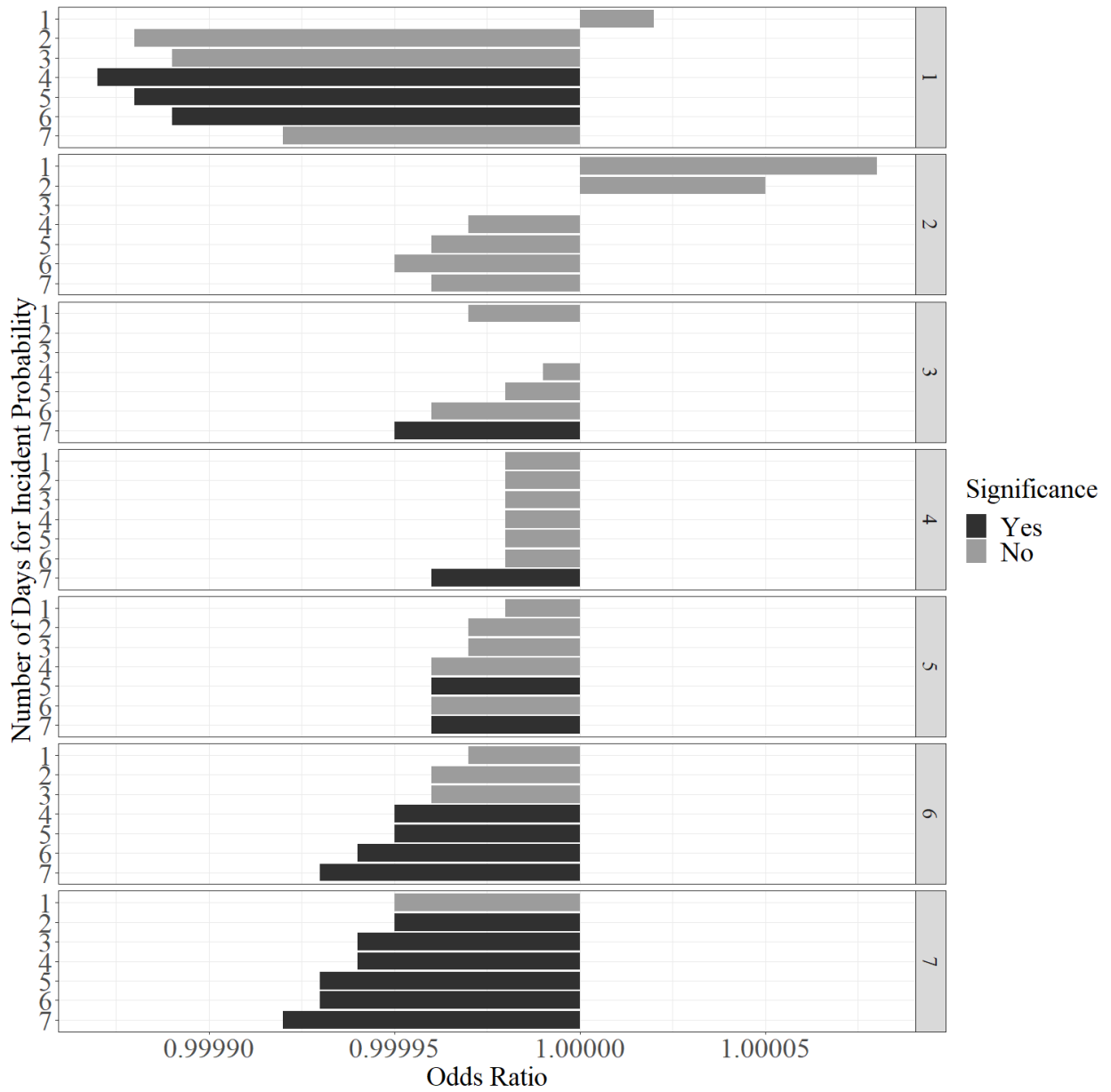


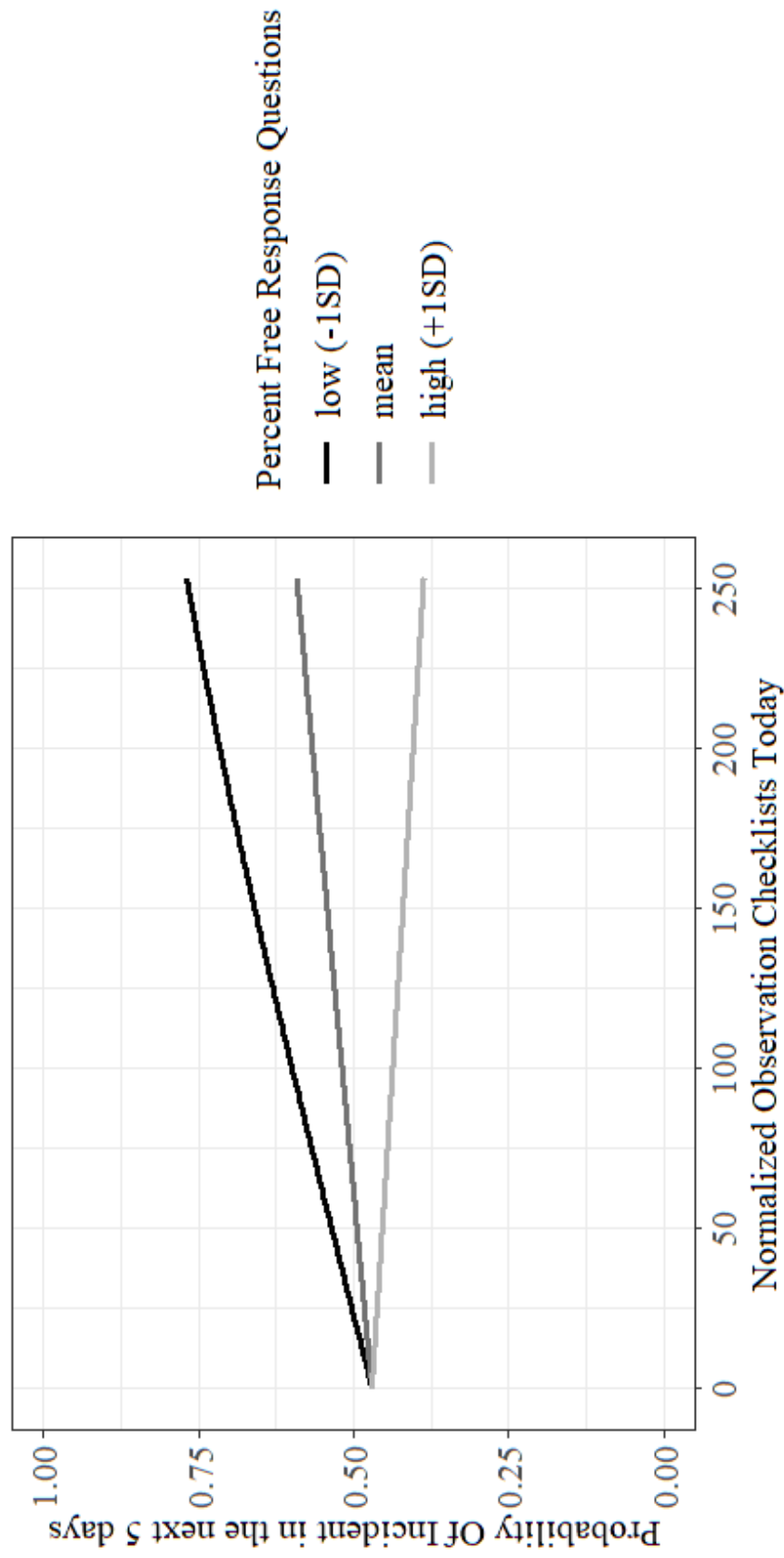
Table 9

*High/Low Interaction Impact on Incident Probabilities
(Free Response Percentage)*

IQV Value	Percentile				
	0th	25th	50th	75th	100th
	0	63	126	190	253
SD - 1	0.470	0.551	0.631	0.705	0.769
SD + 1	0.470	0.449	0.428	0.407	0.387

Figure 11

Checklists with Higher Free Response Predict Lower Incident Probability



Hypothesis 3b: The negative relationship between observation checklist counts and incident probability will be stronger when those observation checklists have free response questions with longer responses.

The one through seven day average lengths of free-response item answers were entered as an interaction term between forty nine logistic regression models assessing the impact of observation sums in the past one to seven days on incident probabilities within the next one to seven days. The LME consisted of observation counts today predicting an incident within the next three days (OR = 1.004, $p = 0.034$), and the overall model for this LME significantly contributed to the reduction in error variance of the outcome from the null model ($R^2_N = 0.0101$, $\chi^2 = 8.06$, $p = 0.045$). The interaction effect for the same LME was nonsignificant (OR < 1.000, $p = .530$), although as shown in Figures 10 and 11, there are clear significant trends across the five, six, and seven day observation count totals on incidents, and the corresponding interaction effects.

Figure 12

*Simple Effect of Observations on Incidents
(Free Response Length)*

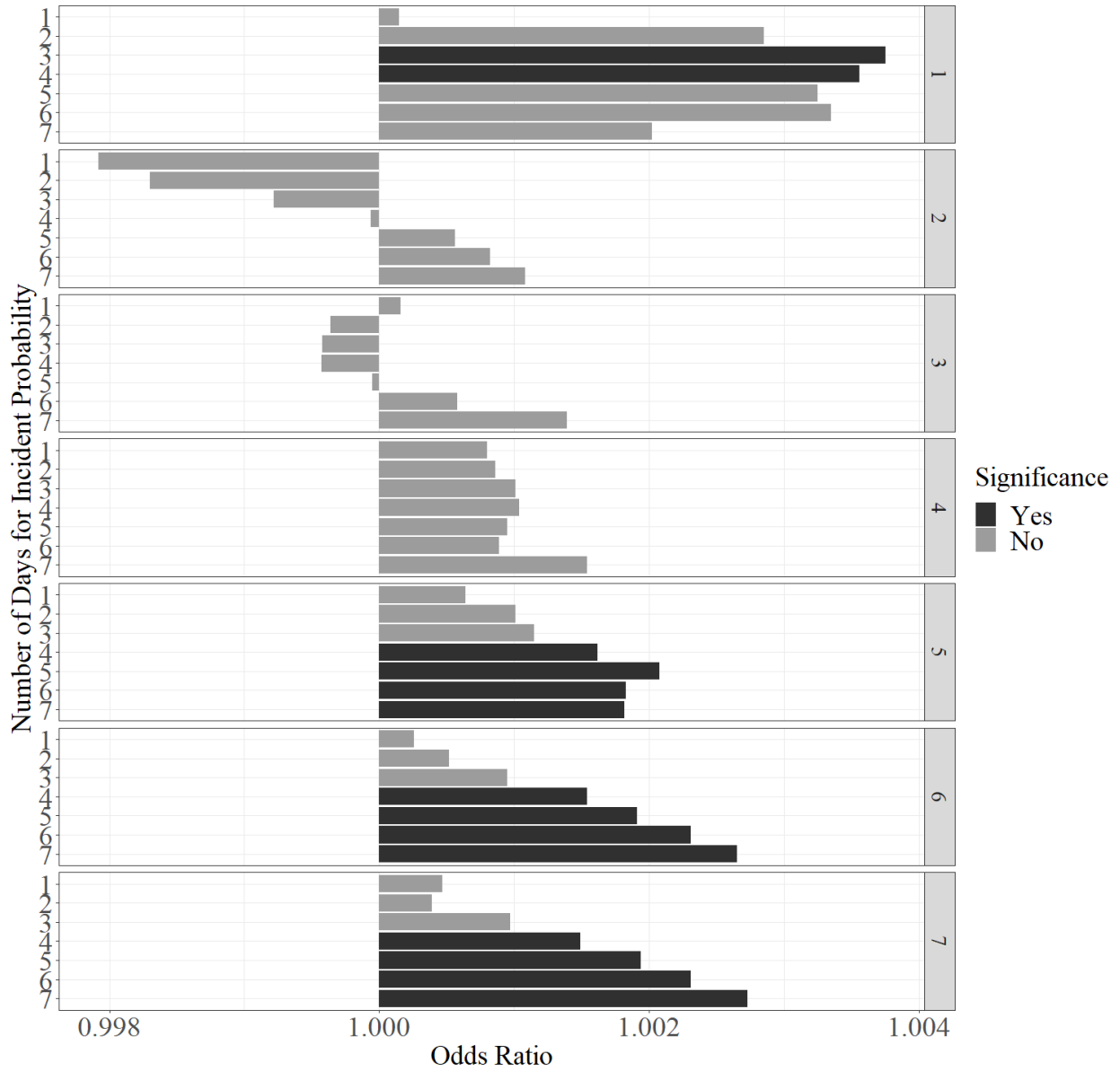
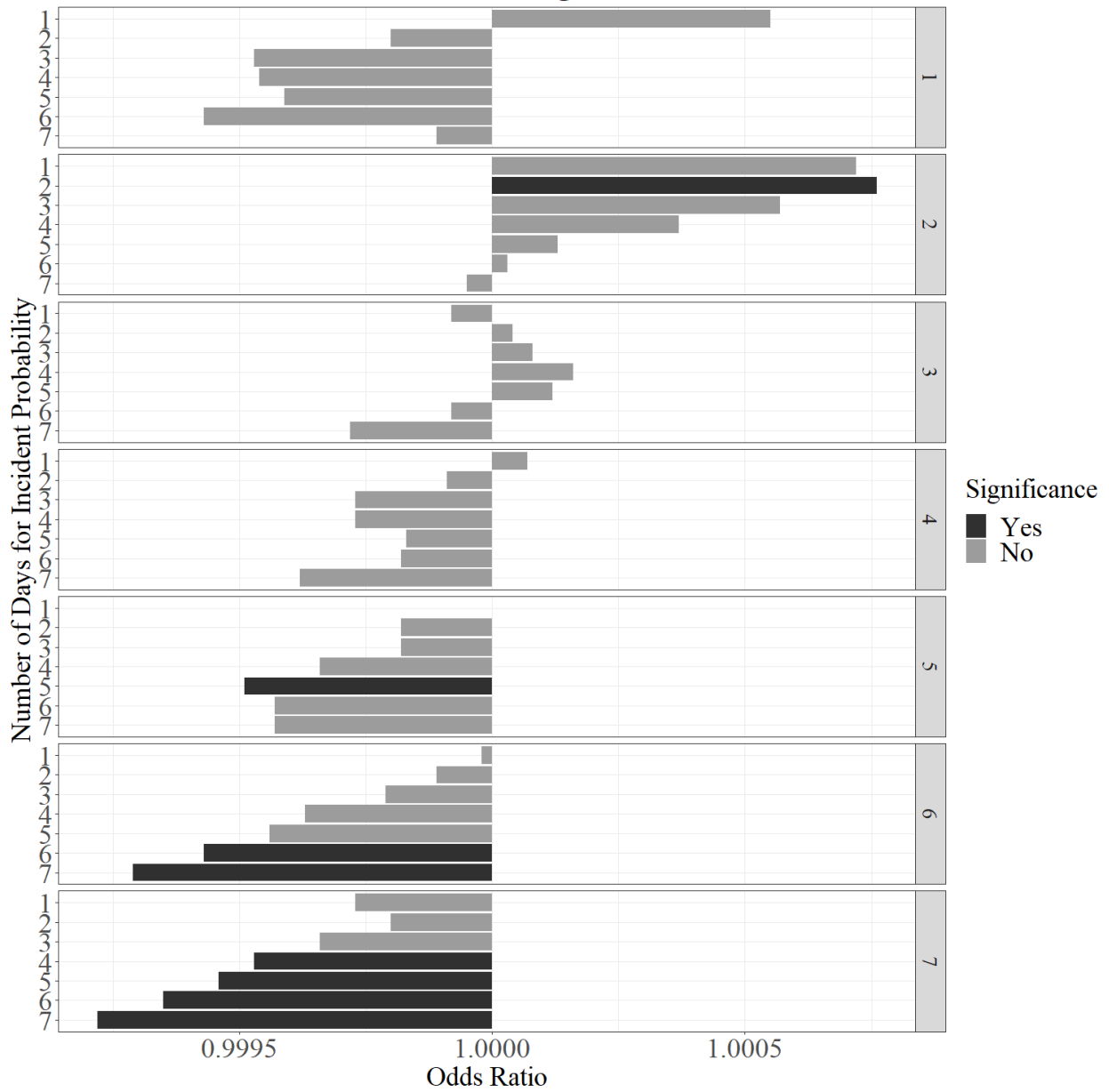


Figure 13

*Odds Ratios of Interactions
between Observation Counts and Answer Length*



Results Summary

Table 10 contains summaries for each of the MLE regressions performed above, including Odds Ratios for all observation and interaction effects, lags, Nagelkerke's pseudo R^2 , likelihood ratio tests, and predictive measures. Additional measures can be assessed in Appendix B, where a greater number of statistics for each test are listed.

Predictably, the model for hypothesis 3 had the greatest area under the curve, highest specificity, and lowest sensitivity, as only 0.5% of days contained incidents, but almost all models were more likely to commit a false-negative error than a false-positive. The only exception to this rule was hypothesis 1, which predicted the seven-day rolling aggregation of incidents. This makes sense, as an incident was more likely than not to occur over any given seven-day period. The most balanced model in regards to specificity and sensitivity was the interaction between free response percentage and observations.

All but Hypothesis 2 were shown to be better at reducing the error variance in the outcome than the null model, as shown by the p-value under 'Overall Model Fit Statistics'.

Table 10*Summary Table for all Effect Sizes and p-Values of Lags with Maximal Effects*

Hypothesis	LME Days Lagged		Observation Effects		Interaction effects	
	Observations	Incidents	OR	p	OR	p
1A	2	7	1.007	0.035	<1.000	0.023
1B	2	2	1.002	0.049	0.966	0.004
2	4	1	1.011	0.032	0.998	0.031
3A	1	5	1.004	0.006	<1.000	0.021
3B	1	3	1.004	0.034	1.000	0.53

Table 10 (continued)*Summary Table for all Effect Sizes and p-Values of Lags with Maximal Effects*

Hypothesis	Overall Model Fit						
	Statistics			Predictive Measures			
	R^2_N	χ^2	p	Accuracy	Specificity	Sensitivity	AUC
1A	0.0098	7.82	0.050	0.640	0.027	0.997	0.525
1B	0.0205	15.2	0.002	0.748	0.998	0.007	0.581
2	0.0559	4.01	0.260	0.994	1.000	0.000	0.748
3A	0.0352	29.3	< 0.001	0.564	0.484	0.641	0.591
3B	0.0101	8.06	0.045	0.651	0.999	0.005	0.556

Discussion

This study sought to add to the BBS literature on observation checklist quality in order to inform both practitioners and researchers on best practices regarding observation checklist creation for incident reduction. The study operationalized quality across checklist design and response variability. Checklist design features include question discrimination score, percentage of safety-confirming questions, percentage of free-response questions. Response quality was defined as fixed-response variability and free-response answer length. Results show that all forms of checklist design are important for enabling observation checklist counts to reduce incident probability.

BBS is only as effective as the data it collects and the behaviors it reinforces. As a system which differentiates itself by its ability to focus resources on achieving specific desirable outcomes rather than preventing a plethora of undesirable ones, the system must be able to inform front-line employees and supervisors on those desired outcomes. Additionally, without quality behavioral information, BBS cannot hope to achieve the metrics required to prevent downstream injuries. The lynchpin to ensuring quality data collection for BBS outcomes is the safety checklist, and this study shows how discrepancies in the quality of checklist design lead to undesirable results.

Two measures of response quality did not significantly affect the relationship between number of observations and incidents. For the first measure, IQV, while the model using information over the past four days to predict incidents tomorrow showed the strongest significant effect between observation counts and incident likelihood, the information provided by the model failed to contribute more meaningfully to incident likelihood than the null model. This finding contrasts to previous literature investigating pencil-whipping as a contributing

factor to ineffective safety processes (Ludwig, 2014). This poor data entry method results in unvaried responses as observers simply check an “all-clear” without truly observing their coworkers’ behaviors. Without the behavioral information necessary for intervention, safety incidents begin to increase. While the current study does not find support for the outcome, it is possible that in this case employees are truly just seeing generally safe results in their environment. Even more likely, the extremely low incident base-rate in the department-level analyses (less than 1%) may have decreased the analyses’ power so significantly that it missed finding a true effect. Future analyses on larger datasets may prove more fruitful for finding an effect.

Second, free-response length did not impact the effect of observation counts on incident counts. Regardless of free-response length, incidents were more likely to occur as employees completed more observation sheets. This baseline positive relationship between checklists and incident probability may result from an overall change in reporting culture across the board. Incident probability may not be increasing because true incidents are occurring more frequently (an increase in true-positives), but that employees who witness an incident are more likely to report that incident (a decrease in false-negatives). Regardless, while free-response length may be an indication of an employee providing a greater amount of detail for a subject matter expert to diagnose potential problems, there is much more to a quality free response question than length, including specificity and discrimination. Future studies attempting to assess free-response quality should pursue these alternatives as potential sources of unstructured information useful to reducing incident likelihood.

Observation checklist design had more impact on incident probability. Discrimination is the first of these operationalizations of quality design to be tested. While both Ludwig and

Johnston & Pennypacker recommend that the pinpoints used to create observation checklists be relatively specific, including information on a specific action verb (“do what?”), a specific physical object being acted upon (“to what?”), a conditional modifier indicating the action’s place within a greater process or purpose (“when?”), and an end-goal or reason for the original action (“why”) (Ludwig, 2018, pp. 113; Johnston & Pennypacker, 1980). Together, these four questions make up what this study calls a Discrimination (What What When Why) Score of checklist item discrimination. A question with a high Discrimination score must have a moderate degree of specificity, as it necessarily describes an action, an object, a context, and a purpose. The creation of specific checklist items aligns with Geller’s recommendations for targeting specific high-risk behaviors in the environment, as greater focus on these behaviors may lead to the largest reduction in safety outcomes (Geller, 2001). In contrast, Komaki et al. suggest using vague observation checklist items, as it allows workers to apply the observation checklist in a broader variety of environments (Komaki et al., 1978; Wirth & Sigurdsson, 2008). The current study supports checklist quality specificity being related to a reduction in incident likelihood. Additionally, BBS relies on reinforcing desired behaviors rather than pointing out undesirable ones. This study tests the veracity of that claim within the context of observation checklists. Observation checklists with a greater average number of safety-confirming items were more likely to prevent future incident occurrence.

Free-response items have been proposed as valuable to safety checklists due to their ability to add contextual information surrounding fixed-response items that might otherwise be overlooked, although this hypothesis had not been tested before this study. The current study supports the use of free-response as a way to collect information, as checklists with a greater percentage of free-response items were more closely related to a reduction in incident likelihood.

Division-wide, the measure of quality with the most immediate impact was safety-confirmation, where safety-confirmative observations over the past two days predicted incident likelihood in the next two days. Safety-confirmative questions reinforce positive behaviors rather than pointing out negative ones, and this study indicates that reinforcement has a short-term impact on incident reduction. The measure slowest to affect incident probability was discrimination, where it took a week before incident likelihood began to decrease. This may be because more specific questions are helpful for subject matter experts reviewing behaviors, who will then have a better understanding of how the employee is deviating from that behavior.

Theoretical Implications

This exploratory analysis of observation quality provides a study design framework for future research. While previous experiments have assessed BBS interventions and have shown a negative relationship between the number of observation checklists and incidents, this study is unique in its exploration into the moderators that drive this relationship (Bogard et al., 2015; Choudhry, 2014; Hagge et al., 2017; Lebbon et al., 2012; Sultzer-Azaroff et al., 1990). The results of this study show that checklist design quality should be considered an important factor when evaluating the effectiveness of interventions in case studies.

As stated previously, this study also sheds light on the debate between generalizability and specificity, where more specific checklist items were related to incident negation. As greater percentages of free-response items was also associated with incident negation, it may also be possible that checklists with more specific action items benefit more greatly from free-response questions, as whatever cannot be captured in the specific fixed-response questions may be captured in the contextual information provided by free-response options. Future studies are needed in order to investigate this interaction.

BBS researchers need to consider both simultaneous and delayed effects of safe behaviors on incidents. Future research should consider more time-series tools to assess the effectiveness of interventions on outcomes. While same-day studies have shown a general decline in safety incidents as employees participate more in observation reporting, the effects of observations on incidents may still be underestimated, as the peak impact of observations may be delayed until a few days later. Ultimately, multivariate time-series models incorporating a variety of safety information may be most helpful toward preventing future incidents.

Practitioner Implications

This study informs checklist design and analyses for practitioners. First, free-response questions (FRQs) are shown to be effective for incident mitigation. Practitioners should ensure they implement a greater number of free-response questions into their checklists for greater contextual factors around specific behaviors. Managers should encourage employees to use the free response prompts, but not necessarily push their employees to write for the sake of writing, as answer length had no impact on observation effectiveness. Additionally, managers should double-check pinpoints to ensure they confirm safe behaviors rather than point out unsafe ones. Feedforward performance appraisal has been shown as effective to ensuring direct reports digest and internalize manager suggestions, and BBS's focus on desirable behaviors and external causes of unsafe behaviors may benefit from the same principle (Budworth et al., 2015).

Finally, as stated among theoretical implications, safety analysts should ensure that they are measuring the impact of interventions at a lag, as same-day evaluations may be fruitless. Instead, practitioners should try to evaluate effectiveness across several different time periods or, barring that, at least select a theoretical appropriate lag between the intervention and outcome.

Limitations and Future Analyses

There are a variety of methodological limitations within this study. Firstly, the findings should be cross-validated on another dataset. This study used an exploratory methodology in which the lag with the strongest relationship between observations and incidents was identified. However, it is possible that these lags are idiosyncratic and therefore do not generalize to different settings. Accordingly, future studies should seek to verify these assumptions on new data.

There was no measure set in place to control for the increase in false-positives resulting from doing so many logistic regressions for each hypothesis. Future analyses may want to lower alpha to 0.01 or lower in order to ensure true effects are found. As this study is exploratory in nature, an alpha of 0.05 was important for finding results for worth pursuing farther.

Four of the five hypotheses were tested across two different divisions within the same organization. Together, these divisions contain upwards of 1,500 employees, and work across many different areas of the plants. While exploratory analysis of intra-division information may offer proof-of-concept for future studies, it is difficult to draw a direct connection between an observation checklist today and an incident tomorrow, due to geographic dispersion of observation and incident locations. Future analyses on the same hypotheses at a lower-level of analysis (I.E. the department level) may improve the predictive validity of the models, since an observation of behavior in any given area is much more likely to cause incidents in the same area. With lower variation within divisions, there should be an increase in power as well. Additional studies in different industries would also lend credence to the generalizability of these results.

In terms of measures, discrimination and safety-confirmation ratings of checklist items was rated by one individual, and so there was no methodological way to account for inter-rater

variability. This methodological shortcoming may reduce the generalizability of item discrimination conclusions for other raters. Future studies should use a variety of raters to assess discrimination categories. In addition, both discrimination (operationalized as Discrimination Score) and safety-confirmation were manually coded. While this method has been shown to be more accurate at identifying free-text data than machine-learning algorithms (van Atteveldt et al. 2021), it lacks scalability. To evaluate the thousands more questions and answers in a dataset, or generate insights quickly from new data, future studies should consider using natural-language processing (NLP) technique to evaluate question quality. NLP is a method that uses machine learning to understand text in a similar way to humans, and has shown success in previous exploratory analyses using free text data from incident reports to predict incident outcomes, injury risk, and injury severity (Sarkar & Maiti, 2020). While these studies have primarily taken a “everything but the kitchen sink” approach to text analysis, creating black-box tools that are accurate but difficult to interpret, a supervised machine learning approach using theory-driven indicators of safety outcomes may generate a model that is both accurate and interpretable. Such a model may be scalable to large amounts of data, and assist practitioners in creating high-quality observation checklists.

References

- Bogard, K., Ludwig, T., Staats, C., & Kretschmer, D. (2015). An Industry's Call to Understand the Contingencies Involved in Process Safety: Normalization of Deviance. *Journal of Organizational Behavior Management*.
- Bradler, C., Dur, R., Neckermann, S., & Non, A. (2016). Employee Recognition and Performance: A Field Experiment. *Management Science*, 62(11), 3085–3099. <https://doi.org/10.1287/mnsc.2015.2291>
- Brondolo, E. (2021). Chapter 16—Methods: Measuring variables. In E. Brondolo (Ed.), *Psychology Research Methods* (pp. 423–449). Academic Press. <https://doi.org/10.1016/B978-0-12-815680-3.00016-5>
- Budworth, M.-H., Latham, G. P., & Manroop, L. (2015). Looking Forward to Performance Improvement: A Field Test of the Feedforward Interview for Performance Management. *Human Resource Management*, 54(1), 45–54. <https://doi.org/10.1002/hrm.21618>
- Choudhry, R. M. (2014). Behavior-based safety on construction sites: A case study. *Accident Analysis & Prevention*, 70, 14–23. <https://doi.org/10.1016/j.aap.2014.03.007>
- Cooper, M. D., (2008). Exploratory Analyses of the Effects of Managerial Support and Feedback Consequences on Behavioral Safety Maintenance. *Journal of Organizational Behavior Management*, 26 (3), 1-41.
- Cooper, M. D. (2009). Behavioral Safety Interventions A Review of Process Design Factors. *Professional Safety*, 54(02).

- DePasquale, J., & Geller, S., (1999). Critical Success Factors for Behavior-Based Safety: A Study of Twenty Industry-wide Applications. *Journal of Safety Research*, 30 (4), 237-249.
- Dufek, J. S., Bates, B. T., & Davis, H. P. (1995). The effect of trial size and variability on statistical power. *Medicine & Science in Sports & Exercise*, 27(2), 288–295.
<https://doi.org/10.1249/00005768-199502000-00021>
- Fudenberg, D., & Levine, D. K. (2014). Recency, consistent learning, and Nash equilibrium. *Proceedings of the National Academy of Sciences*, 111(Supplement 3), 10826.
<https://doi.org/10.1073/pnas.1400987111>
- Geller, E.S. (2001) *The Psychology of Safety Handbook*. CRC Press.
- Geller, E. S. (2005). Behavior-Based Safety and Occupational Risk Management. *Behavior Modification*, 29(3). <https://doi.org/10.1177/0145445504273287>
- Heinrich, H. (1931). *Industrial Accident Prevention; a Scientific Approach*. McGraw-Hill book company.
- Hagge, M., McGee, H., Matthews, G., & Aberle, S. (2017). Behavior-Based Safety in a Coal Mine: The Relationship Between Observations, Participation, and Injuries Over a 14-Year Period. *Journal of Organizational Behavior Management*, 37(1), 107-118.
- Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research*. Lawrence Erlbaum Associates, Inc.
- Kabil, G. V., & Sundararaju, V. (2019). Behavior Based Safety in Workplace. *International Journal of Research in Engineering, Science, and Management*, 2(12).
- Kluger, A. N., & Nir, D. (2010). The feedforward interview. *Human Resource Management in Israel*, 20(3), 235–246. <https://doi.org/10.1016/j.hrmr.2009.08.002>

- Komaki, J., Barwick, K. D., & Scott, L. R. (1978). A behavioral approach to occupational safety: Pinpointing and reinforcing safe performance in a food manufacturing plant. *Journal of Applied Psychology*, 63(4), 434–445. <https://doi.org/10.1037/0021-9010.63.4.434>
- Lebbon, A., Sigurdsson, S., & Austin, J. (2012). Behavioral Safety in the Food Services Industry: Challenges and Outcomes. *Journal of Organizational Behavior Management*, 32 (1). <https://doi.org/10.1080/01608061.2011.592792>.
- Laske, M. (2020). Who Is Better At Identifying At-Risk Behavior? Leader Versus Employee Processes To Implement Task-Specific Behavioral Pinpoints. *Unpublished Master's Thesis*. Appalachian State University, Boone, NC.
- Ludwig, T. D. (2014). The Anatomy of Pencil Whipping. *Professional Safety*, 59(2), 47–50.
- Ludwig, T. D. (2018). Dysfunctional practices that kill your safety culture. Calloway Publishing.
- McSween, T. (2003). Value-Based Safety Process: Improving Your Safety Culture with Behavior-Based Safety, Second Edition. John Wiley & Sons, Inc., Kindle Version.
- Mayer, G. R., Sultzer-Azaroff, B., & Wallace, M. (2019). Behavior analysis for lasting change (4th ed.). Sloan Publishing, LLC.
- Miller, L. K. (2006). Principles of everyday behavior analysis (4th ed.). Thomson Wadsworth.
- National Safety Council. (2020) *Work Injury Costs*. <https://injuryfacts.nsc.org/work/costs/work-injury-costs/>
- Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting Final Course Performance From Students' Written Self-Introductions: A LIWC Analysis. *Journal of Language and Social Psychology*, 32(4), 469–479. <https://doi.org/10.1177/0261927X13476869>

- Sarkar, S., & Maiti, J. (2020). Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. *Safety Science*, 131, 104900. <https://doi.org/10.1016/j.ssci.2020.104900>
- Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. (p. 457). Appleton-Century.
- Speer, A. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, 71. <https://doi.org/10.1111/peps.12263>
- Speer, A. B. (2021). Scoring Dimension-Level Job Performance from Narrative Comments: Validity and Generalizability When Using Natural Language Processing. *Organizational Research Methods*, 24(3), 572–594. <https://doi.org/10.1177/1094428120930815>
- Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review*, 21(2), 85–95. <https://doi.org/10.1016/j.hrmr.2010.09.013>
- Sulzer Azaroff, B. (1980). Behavioral Ecology and Accident Prevention. *Journal of Organizational Behavior Management*, 2(1), 11–44. https://doi.org/10.1300/J075v02n01_02
- Sultzer-Azaroff, B., Loafman, B., Merante, R., & Hlavacek, A. (1990). Improving Occupational Safety in a Large Industrial Plant. *Journal of Organizational Behavior Management*, 11(1), 99-120. https://doi.org/10.1300/J075v11n01_07.
- Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie Canadienne de Psychiatrie de l'enfant et de l'adolescent*, 19(3), 227–229. PubMed.

U.S Bureau of Labor Statistics. (2020). *BLS Census of Fatal Occupational Injuries Summary, 2019*. <https://www.bls.gov/news.release/cfoi.nr0.htm>

U.S. Bureau of Labor Statistics. (2021). Employer-Reported Workplace INjuries and Illnesses - 2020.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2), 121–140. <https://doi.org/10.1080/19312458.2020.1869198>

Wirth, O., & Sigurdsson, S. O. (2008). When workplace safety depends on behavior change: Topics for behavioral safety research. *Journal of Safety Research*, 39(6), 589–598. <https://doi.org/10.1016/j.jsr.2008.10.005>

Appendix A: Logistic Regression Results

Hypothesis 1A

Table 11

Model Coefficients - 7-Day Incident Aggregation

Variable	Estimate	95% Confidence Interval	
		Lower	Upper
Intercept	-0.64231	-1.50469	0.2201
2-Day Observation Sum	0.00566	3.90E-04	0.0109
2-Day Actionability Score Average	0.01792	0.00493	0.0309
2-Day Observation Sum * 2-Day Actionability Score Average	-8.15e-5	-1.52e-4	-1.10e-5

Table 11
(continued)

Model Coefficients - 7-Day Incident Aggregation

	SE	Z	p	Odds Ratio	95% Confidence Interval	
					Lower	Upper
Intercept	0.44	-1.46	0.144	0.526	0.222	1.246
2-Day Observation Sum	0.00269	2.1	0.035	1.006	1	1.011
2-Day Actionability Score Average	0.00663	2.7	0.007	1.018	1.005	1.031
2-Day Observation Sum * 2-Day Actionability Score Average	3.60E-05	-2.27	0.023	1	1	1

Table 12*Model Fit Measures*

Model	Deviance	AIC	R ² N	Overall Model Test		
				χ^2	df	p
1	1427	1435	0.00976	7.82	3	0.05

Table 13*Collinearity Statistics*

Variable	VIF	Tolerance
2-Day Observation Sum	13.46	0.0743
2-Day Actionability Score Average	4.46	0.2241
2-Day Observation Sum * 2-Day Actionability Score Average	21.96	0.0455

Table 14*Classification Table – 7-Day Incident Aggregation*

Observed	Predicted		% Correct
	0	1	
0	11	390	2.74
1	2	688	99.7

Table 15*Predictive Measures*

Accuracy	Specificity	Sensitivity	AUC
0.641	0.0274	0.997	0.525

Table 16

Omnibus Likelihood Ratio Tests

Predictor	χ^2	df	p
2-Day Observation Sum	4.47	1	0.034
2-Day Actionability Score Average	7.55	1	0.006
2-Day Observation Sum * 2- Day Actionability Score Average	5.18	1	0.023

Hypothesis 1B

Table 17

Model Coefficients - 7-Day Incident Aggregation

Variable	Estimate	95% Confidence Interval		SE
		Lower	Upper	
Intercept	-1.72771	-2.23143	-1.22398	0.25701
2-Day Observation Sum	0.00228	5.09E-06	0.00456	0.00116
2-Day Average Affirmation Percent	1.02989	0.48129	1.5785	0.2799
2-Day Observation Sum * 2-Day Average Affirmation Percent	-0.00439	-0.00736	-0.00142	0.00152

Table 17 (continued)

Model Coefficients - 7-Day Incident Aggregation

Variable	Z	p	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Intercept	-6.72	< .001	0.178	0.107	0.294
2-Day Observation Sum	1.96	0.049	1.002	1	1.005
2-Day Average Affirmation Percent	3.68	< .001	2.801	1.618	4.848
2-Day Observation Sum * 2-Day Average Affirmation Percent	-2.9	0.004	0.996	0.993	0.999

Table 18*Model Fit Measures*

Model	Deviance	AIC	R ² N	Overall Model Test		
				χ^2	df	p
1	1217	1225	0.0205	15.2	3	0.002

Table 19*Collinearity Statistics*

Variable	VIF	Tolerance
2-Day Observation Sum	2.04	0.49
2-Day Average Affirmation Percent	6.23	0.161
2-Day Observation Sum * 2-Day Average Affirmation Percent	5.5	0.182

Table 20*Classification Table – 7-Day Incident Aggregation*

Observed	Predicted		% Correct
	0	1	
0	814	2	99.8
1	273	2	0.727

Table 21*Predictive Measures*

Accuracy	Specificity	Sensitivity	AUC
0.748	0.998	0.00727	0.581

Table 22

Omnibus Likelihood Ratio
Tests

Predictor	χ^2	df	p
2-Day Observation Sum	3.81	1	0.051
2-Day Average Affirmation Percent	13.65	1	< .001
2-Day Observation Sum * 2- Day Average Affirmation Percent	8.65	1	0.003

Hypothesis 2

Table 23

Model Coefficients - 7-Day Incident Aggregation

Variable	Estimate	95% Confidence Interval		SE
		Lower	Upper	
Intercept	-12.279	-20.2	-4.3424	4.04938
4-Day Observation Sum	0.0111	9.46E-04	0.0212	0.00518
4-Day IQV Average	0.1422	-7.36e-4	0.2851	0.07292
4-Day Observation Sum * 4-Day IQV Average	-2.27e-4	-4.33e-4	-2.04e-5	1.05E-04

Table 23 (continued)

Model Coefficients - 7-Day Incident Aggregation

Variable	Z	p	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Intercept	-3.03	0.002	4.65E-06	1.66E-09	0.013
4-Day Observation Sum	2.14	0.032	1.011	1.001	1.0215
4-Day IQV Average	1.95	0.051	1.153	0.999	1.3299
4-Day Observation Sum * 4-Day IQV Average	-2.15	0.031	1	1	1

Table 24*Model Fit Measures*

Model	Deviance	AIC	R ² N	Overall Model Test		
				χ^2	df	p
1	70.2	78.2	0.0559	4.01	3	0.26

Table 25*Collinearity Statistics*

Variable	VIF	Tolerance
4-Day Observation Sum	15.29	0.0654
4-Day IQV Average	4.34	0.2304
4-Day Observation Sum * 4-Day IQV Average	11.51	0.0869

Table 26*Classification Table – 7-Day Incident Aggregation*

Observed	Predicted		% Correct
	0	1	
0	1069	0	100
1	6	0	0

Table 27*Predictive Measures*

Accuracy	Specificity	Sensitivity	AUC
0.994	1	0	0.748

Table 28Omnibus Likelihood Ratio
Tests

Predictor	χ^2	df	p
4-Day Observation Sum	3.29	1	0.07
4-Day Average Index of Qualitative Variability	3.83	1	0.05
4-Day Observation Sum * 4- Day Average Index of Qualitative Variability	3.59	1	0.058

Hypothesis 3A

Table 29

Model Coefficients - 7-Day Incident Aggregation

Variable	Estimate	95% Confidence Interval		SE
		Lower	Upper	
Intercept	-0.12089	-0.39784	0.15606	0.1413
1-Day Observation Sum	0.00416	0.00122	0.00711	0.0015
One-Day Free Response Percentage	5.04E-05	-0.01015	0.01025	0.00521
1-Day Observation Sum * One-Day Free Response Percentage	-1.21e-4	-2.24e-4	-1.83e-5	5.26E-05

Table 29 (continued)

Model Coefficients - 7-Day Incident Aggregation

Variable	Z	p	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Intercept	-0.85555	0.392	0.886	0.672	1.169
1-Day Observation Sum	2.77381	0.006	1.004	1.001	1.007
One-Day Free Response Percentage	0.00968	0.992	1	0.99	1.01
1-Day Observation Sum * One-Day Free Response Percentage	-2.30792	0.021	1	1	1

Table 30*Model Fit Measures*

Model	Deviance	AIC	R ² N	Overall Model Test		
				χ^2	df	p
1	1486	1494	0.0352	29.3	3	< .001

Table 31*Collinearity Statistics*

Variable	VIF	Tolerance
1-Day Observation Sum	1.45	0.692
One-Day Free Response Percentage	5.14	0.195
1-Day Observation Sum * One-Day Free Response Percentage	5.91	0.169

Table 32*Classification Table – 7-Day Incident Aggregation*

Observed	Predicted		% Correct
	0	1	
0	261	278	48.4
1	199	355	64.1

Table 33*Predictive Measures*

Accuracy	Specificity	Sensitivity	AUC
0.564	0.484	0.641	0.591

Table 34

Omnibus Likelihood Ratio Tests

Predictor	χ^2	df	p
1-Day Observation Sum	7.85	1	0.005
One-Day Free Response Percentage	9.37E-05	1	0.992
1-Day Observation Sum * One-Day Free Response Percentage	5.35	1	0.021

Hypothesis 3B

Table 35

Model Coefficients - 7-Day Incident Aggregation

Variable	Estimate	95% Confidence Interval		SE
		Lower	Upper	
Intercept	-0.87712	-1.23159	-0.5227	0.18085
One-Day Observation Sum	0.00666	0.00212	0.0112	0.00231
Average FRQ Response Length	-4.02e-5	-1.11e-4	3.03E-05	3.60E-05
One-Day Observation Sum * Average FRQ Response Length	-1.67e-7	-6.56e-7	3.22E-07	2.49E-07

Table 35 (continued)

Model Coefficients - 7-Day Incident Aggregation

Variable	Z	p	Odds Ratio	95% Confidence Interval	
				Lower	Upper
Intercept	-4.85	< .001	0.416	0.292	0.593
One-Day Observation Sum	2.876	0.004	1.007	1.002	1.011
Average FRQ Response Length	-1.117	0.264	1	1	1
One-Day Observation Sum * Average FRQ Response Length	-0.67	0.503	1	1	1

Table 36*Model Fit Measures*

Model	Deviance	AIC	R ² N	Overall Model Test		
				χ^2	df	p
1	1403	1411	0.0141	11.2	3	0.011

Table 37*Collinearity Statistics*

Variable	VIF	Tolerance
One-Day Observation Sum	3.52	0.2843
Average FRQ Response Length	6.49	0.154
One-Day Observation Sum * Average FRQ Response Length	10.17	0.0983

Table 38*Classification Table – 7-Day Incident Aggregation*

Observed	Predicted		% Correct
	0	1	
0	704	7	99
1	382	0	0

Table 39*Predictive Measures*

Accuracy	Specificity	Sensitivity	AUC
0.644	0.99	0	0.569

Table 40

Omnibus Likelihood Ratio Tests

Predictor	χ^2	df	p
One-Day Observation Sum	8.458	1	0.004
Average FRQ Response Length	1.246	1	0.264
One-Day Observation Sum * Average FRQ Response Length	0.456	1	0.5

Appendix B: IRB Approval Form



INSTITUTIONAL REVIEW BOARD
Office of Research Protections
ASU Box 32068
Boone, NC 28608
828.262.2692
Web site: <http://researchprotections.appstate.edu>
Email: irb@appstate.edu
Federalwide Assurance (FWA) #00001076

To: Timothy Ludwig
Psychology , Psychology
CAMPUS EMAIL

From: IRB Administration

Date: 9/28/2020

RE: Determination that Research or Research-Like Activity does not require IRB Approval

Agrants #: 18-0155

Grant Title: Sponsors: National Institute for Occupational Safety and Health (NIOSH)

STUDY #: 19-0072

STUDY TITLE: Developing an Analytics Strategy to Describe, Diagnose, and Predict Workplace Safety Outcomes

The IRB determined that the activity described in the study materials does not constitute human subject research as defined by University policy and the federal regulations [45 CFR 46.102 (e or l)] and does not require IRB approval.

Study Regulatory and other findings:

As described in the application and emphasized in the responses to clarifications requested by the reviewer, this project is best understood as program evaluation and quality improvement and thus does not meet the definition of research as defined in 45 CFR 46.102(l).

This determination may no longer apply if the activity changes. IRB approval must be sought and obtained for any research with human participants.

If you have any questions about this determination, please contact the IRB Administration at 828-262-4060 or irb@appstate.edu.

Thank you.

CC:

Yalcin Acikgoz, Psychology
Shawn Bergman, Psychology
Julie Brooks, Psychology
Cori Ferguson, Psychology
Nicholas Granowsky, Psychology
Ava Young, Psychology

Appendix C: Data and Materials Distribution Agreement

Safety Analytics Project Data and Materials Distribution Agreement

This Data and Materials Distribution Agreement (hereinafter "Agreement"), effective as of **March 26, 2018 ("Effective Date")**, by and between **Eastman Chemical Company** (hereinafter referred to as "Eastman"), and **Appalachian State University** (hereinafter referred to as "Appalachian State"), an educational institution of the State of North Carolina, on behalf of **Drs. Shawn Bergman and Timothy Ludwig** (hereinafter referred to as the "Principal Investigators" of the propose research project).

The undersigned parties hereby enter into this Data and Materials Distribution Agreement (DMDA) as of the date specified on the final page hereof.

DEFINITIONS

For purposes of this agreement, the terms below have the following meaning:

"**Data**" refers to any and all study data, information, and associated records obtained directly from Eastman.

"**Personally Identifying Information**" refers to information that could be used to identify an individual's identity.

"**Materials**" refers to Eastman information and materials, including but not limited to HR personnel data (e.g., selection tests, hiring decisions, and performance management data) provided to the Recipient.

"**Resultant Data**" refers to data derived in whole or in part by Recipient from Data and/or Materials provided under this DMDA.

"**Eastman Personnel**" refers to the personnel of the Eastman authorized to provide the University with data and materials and who are authorized review and provide feedback on the presentation and/or publication of project's results.

"**Research Project**" refers to the project described in section 4 below.

"**Recipient**" refers to the research investigators at Appalachian State receiving access to the Eastman Study Data and/or Materials requested for the Research Project identified in section 3 below.

"**Principal Investigator (PI)**" refers to the Research Project directors for the Recipient.

TERMS AND CONDITIONS

It is mutually agreed as follows:

1. **Materials.** Eastman, in its sole discretion, agrees to transfer to Recipient the Materials described as follows, for use by the Recipient's PI to conduct the Research Project as summarized in section 4 below:
 - a. **Not applicable.**
2. **Data.** Eastman, in its sole discretion, agrees to provide Recipient with Data described as follows: Safety data from Eastman's safety data collection systems with PII redacted.

3. Notwithstanding Paragraph 15, below, Eastman may terminate its agreement to provide the above Materials and Data at any time with 30 days prior notice to Appalachian State.

4. Research Project.

- 4.1 These Materials and Data will be used by Recipient's PI solely in connection with the Research Project, as named and described below:

Project title: Safety Analytics Data Audit and Phase I Investigation

Project description: Eastman's HSE program assesses a number of factors related to environmental and worker safety. While this information has been useful in helping Eastman monitor and manage safety risks, full potential of the information collected may not be realized. More specifically, data analytics can use the information collected by Eastman to identify the most effective response to unplanned events and leading indicators to direct resources that optimally mitigate risks in the future. The project is to investigate the efficacy of data analytics to achieve these results.

- 4.2 This DMDA covers only the Research Project cited in section 4.1 of this DMDA. Recipient must submit a separate DMDA for each Research Project for which Data and/or Materials are requested.

5. Non-transferability. This DMDA is not transferable.

- 5.1 Recipient and Recipient's PI agree that substantive changes made to the Research Project, and/or appointment by Recipient of another Principal Investigator and/or transfer of Recipient's PI to another institution or other entity to complete the Research Project, require execution of a separate DMDA. Recipient may not distribute Data or Materials to any other individual or entity, regardless of the intended use of such Data or Materials. However, nothing in this section precludes Recipient from publishing results of the Research Project through the usual channels of scientific publication provided that the appropriate approvals have been given by Eastman Personnel, see section 7 below.

6. Conduct of Research Project. Recipient's PI is responsible for the conduct of the Research Project and shall be responsible for assuring that any co-investigator(s) comply with the terms of this DMDA.

7. Publication. Publication and presentation of the results of the Research Project is allowed.

- 7.1 The Recipient's PI is to provide the authorized representative for the Eastman a copy of submission materials ten (10) business days in advance of submission for presentation, oral and written presentation materials (e.g., study write-up, visual aids, and/or slides), fifteen (15) business days in advance of presenting Research Project findings, and any manuscript or other disclosure document twenty (20) business days in advance of submission for publication, in order to permit review and provide comments and feedback, and to ensure compliance with the requirements of this DMDA.
- 7.2 At the request of Eastman, the source of the data will be anonymized so the source of the data presented in the publication and presentation is not revealed.
- 7.3 At the request of Eastman, data presented in any publication and presentation will be standardized (e.g., data will be presented in terms of z-scores or mean-centered)

so that averages are presented as zeros, the absolute levels of the values of the data cannot be determined, and the relationships between variables and concepts are retained.

7.4 The Recipient may not submit or present any results of the Research Project attributed to Eastman without allowing Eastman review and provide comments and feedback by Eastman Personnel.

8. Non-Identification. Recipient and Recipient's PI agree that Materials and/or Data will not be used, either alone or in conjunction with any other information, in any effort to determine the individual the PII of any of the participants from whom Data and/or Materials were obtained or derived.

8.1 Eastman will not share Personally Identifying Information ("PII") with the Recipient. Data provided to the Recipient and Recipient's PI will have PII removed.

8.2 In the event that any PII is discovered in the Data or Materials, the Recipient and Recipient's PI will remove the PII information from the Data or Materials and report the issue immediately to Eastman Personnel.

9. Use Limited to Research Project. Recipient and Recipient's PI agree that Materials, their progeny, or derivatives thereof, and Resultant Data will not be used in any experiments or procedures unless said experiments or procedures are disclosed and approved as part of the Research Project.

10. No Distribution. Recipient and Recipient's PI agree to retain control over Data, Materials and their progeny, and derivatives thereof. To the extent permitted by North Carolina law, Recipient and Recipient's PI further agree not to transfer Data, Materials and their progeny, and derivatives thereof, with or without charge, to any other entity or individual, unless required to disclose pursuant to law, regulation, public records act request, subpoena or court order.

11. Resultant Data to be Provided to Eastman. Recipient and Recipient's PI agree to provide Eastman with a report at least every twelve (12) months during the term of this DMDA. The report shall include a description of the activities performed and Resultant Data obtained during the twelve (12) months before the reporting date. Recipient and Recipient's PI agree that Eastman may distribute all such Resultant Data to requesting parties.

12. Recipient's Compliance with Recipient IRB's Requirements. Recipient certifies that the conditions for use of the Data and/or Materials in conjunction with the Research Project have been reviewed by the Recipient's Institutional Review Board (IRB). Recipient also agrees to report to Recipient's IRB any unanticipated problems or changes in the Research Project that involve additional risks to participants or others. Recipient remains subject to applicable state and local laws and regulations and institutional policies that provide additional protections for human subjects.

13. Recipient's Responsibility to follow Data Security Best Practices. Recipient is aware of computer and data security best practices and will follow them for receipt, storage and use of Data and Resultant Data. An example of Appalachian State's data Secure Storage and Sharing guidelines can be found at:


http://security.appstate.edu/resources/policies_and_standards/data_class_guideline

14. **Amendments.** Amendments to this DMDA must be made in writing and signed by authorized representatives of all parties.
15. **Termination.** This DMDA shall terminate at the earliest of: the completion of the Research Project; five (5) years after the effective date of this DMDA; abandonment of the Research Project; or violation by Recipient of any provisions of this DMDA not remedied within 30 days after the date of notice by Eastman of such violation.
16. **Disqualification, Governing Law.** Failure to comply with any of the terms of this DMDA may result in disqualification of Recipient from receiving additional Data and/or Materials. This Agreement shall be governed by the laws of the State of North Carolina.
17. **Prior Distribution Agreements.** By execution of this DMDA, Recipient certifies its good faith belief that it is in compliance with the terms and conditions of all its existing DMDAs with Eastman.

Appalachian State University

By: 
Name: Charna K. Howson
Title: Director, Sponsored Programs
Date: April 27, 2018

Eastman Chemical Company

By: 
Name: Mark Paul
Title: Director, Global BSES
Date: 5/7/18

Vita

Charles Riggs Matthews was born in Augusta, Georgia, to Kit and Charles Matthews. He graduated from Aiken High School in 2013 before graduating Clemson University with a dual degree in Biological Sciences and Psychology in 2017. He then enrolled in Appalachian State University to study a Master of Arts degree in the Industrial/Organizational Psychology and Human Resource Management program in Boone, NC.