## EVM interview protocol

### Introductions

The interview began with a brief round of introductions. We mostly used this time to build rapport and to give participants the opportunity to ask questions about the logistics and goals of the study. We also confirmed at this time that we had consent to record our interviews, which were conducted on Zoom.

Participants were provided with links to both (1) the EVM interface ([url redacted]) and (2) a Google Doc containing descriptions of the variables in each of the two datasets used in the think aloud portion of the interview. We include the dataset information provided to participants here for completeness:

#### **Forest fires**

1. month - month of the year when the fire happened, 'jan' to 'dec'
2. day - day of the week when the fire happened, 'mon' to 'sun'
3. temp - temperature in Celsius degrees at the location of the fire, 2.2 to 33.30
4. wind - wind speed in km/h at the location of the fire, 0.40 to 9.40
5. rain - whether it rained outside at the location of the fire, ‘no’ or ‘yes’
6. rh - relative humidity in % as the location of the fire, 15 to 100
7. ffmc - Fine Fuel Moisture Code (FFMC) index, a numeric rating of the moisture of material littering the forest floor at the location of the fire, 18.7 to 96.20
8. dmc - Duff Moisture Code (DMC), a numeric rating of the average moisture content of loosely compacted organic layers of moderate underground depth at the location of the fire, 1.1 to 291.3
9. dc - Drought Code (DC) index, a numeric rating of the average moisture content of deep, compact organic underground layers at the location of the fire, 7.9 to 860.6
10. area - the burned area of the forest in hectares (hundreds of acres), 0.00 to 1090.84

#### **Student absences**

1. age - student's age, 15 to 22
2. address - student's home address type, ‘urban’ or ‘rural’
3. internet - internet access at home, ‘no’ or ‘yes’
4. g\_edu - guardian’s level of education, 0 - none, 1 - kindergarten to 4th grade, 2 - 5th to 9th grade, 3 - 10th to 12th grade, 4 - higher education
5. g\_job - guardian’s job, 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home', or 'other'
6. travel\_time - home to school travel time in minutes, from 1.1 to 82.3
7. study\_time - weekly study time in hours, from 0.0 to 14.7
8. failures - number of past classes failed, 0 to 4
9. alcohol - average daily self reported alcohol consumption, from 1 - very low to 5 - very high
10. absences - number of school absences, from 0 to 93

### Demonstration 1

We preceded the first think-aloud session with a brief introduction of EVM and its basic operations without model checking functionality enabled. We used the cars dataset for the demonstration. We first showed participants how to create a basic chart using chart construction, a strip plot of miles per gallon (mpg) grouped by the number of cylinders in a car’s engine (cyl). Then we pointed out that there are observations where values of mpg and cyl equal zero and that these might reflect an error in the dataset. Second, we demonstrated how to use EVM’s filtering functionality to remove these values from the displayed dataset. We also showed how to clear or remove these filters after defining them. Last, we demonstrated how to use EVM’s data transformation functionality to log transform values of mpg.

Participants were given the opportunity to ask clarifying questions about the interface and its features before moving on.

### Think-aloud session 1

At the beginning of think-aloud sessions, participants were asked to share their screen, input a unique identifier to the EVM interface, and were instructed to select either the forest fires or student absences dataset from a dropdown menu. Unique identifiers were used for basic interaction logging purposes, although we were not able to use these data to inform our analysis due to problems with data loss. Participants were randomly assigned to datasets based on a counterbalancing procedure, which was designed to guarantee that each dataset was used an equal number of times with each interface condition (i.e., model checking disabled vs enabled; think-aloud session 1 vs session 2) across participants.

Before participants began exploring the selected dataset, the interviewer read the following prompt verbatim:

“Please spend about 25 minutes exploring the dataset. Try to make sense of it and look for patterns as you might when exploring a new dataset for work. **Your task** is to look into potential influences on *[area burned in forest fires/student absences*]. Imagine that we will collect any observations you feel are worth reporting and send them to a colleague who will take a closer look at the data. While you do this, please **think-aloud** about what you are doing, what you are looking for, and what it makes you think about the data.”

Reported observations and patterns were not collected or analyzed in any way. This prompt was provided with the intention of placing users in a controlled but realistic data exploration scenario where there was a clear outcome variable they were trying to explain. Participants were reminded to think-aloud during the use session. The interviewer also answered direct questions from participants but otherwise avoided interjecting.

### Demonstration 2

After the first think-aloud session ended, the interviewer asked the participants to stop sharing their screen and switched to sharing their own screen. To prepare participants for the second think-aloud session, the interviewer gave a brief introduction of EVM’s model checking functionality. Again, we used the cars dataset for the demonstration. We picked up where demonstration 1 left off with a strip plot comparing between miles per gallon (mpg) and the number of cylinders in a car’s engine (cyl). We started by demonstrating a basic intercept model to answer the question, “Is it possible that the relationship between of mpg and cyl is just noise?” The model specification was mpg ~ 1, ~1 and assumed normally distributed residuals and homogeneity of variance. After fitting this model and seeing its predictions grafted onto the strip plot we authored previously, we pointed out that the pattern in the observed data did not match the pattern in the model predictions. Second, we demonstrated a model comparison attempting to see whether including the variables shown on the chart (i.e, cyl in this case) as predictors in the model would help to improve model fit. The model specification was mpg ~ cyl, ~cyl and assumed normally distributed residuals and heterogeneity of variance across levels of cyl. After fitting this second model, we pointed out that the pattern of predictions more closely matched the observed data than the previous model. We removed the previous intercept model, leaving only the second model we fit in the display. Third, we demonstrated how model check visualizations can be used to try to explain patterns on a chart in terms of variables not shown on the chart. To do this, we removed cyl from our strip plot and replaced it with the variable horsepower hp, creating a scatterplot of mpg and hp. When we did this, EVM continued showing predictions from our second model defined above (i.e., mpg ~ cyl, ~cyl). We pointed out how using only cyl but not hp as a predictor actually does a pretty decent job of capturing the non-linear trade-off between of mpg and hp. Note that these three examples mirror the use cases described in Section 4.1 of the paper.

Again, participants were given the opportunity to ask clarifying questions about the interface and its features before moving on.

### Think-aloud session 2

Again, participants were asked to share their screen, input a unique identifier to the EVM interface, and were instructed to select either the forest fires or student absences dataset from a dropdown menu. Following our counterbalancing procedure, participants who explored one dataset in the first think-aloud session were asked to select the other dataset for the second session.

Before participants began exploring the second dataset, the interviewer read the following prompt verbatim:

“Please spend about 25 minutes exploring this new dataset, now with modeling enabled. Again, try to make sense of the data and look for patterns as you might when exploring a new dataset for work. *Use the model bar when you feel inclined to do so.* **Your task** is to look into potential influences on *[area burned in forest fires/student absences*]. Imagine that we will collect any observations you feel are worth reporting and send them to a colleague who will take a closer look at the data. Please remember to **think-aloud** about what you are doing, what you are looking for, and what it makes you think about the data.”

The second think-aloud proceeded exactly as the first, except with model checking functionality now enabled in EVM.

### Debriefing interview

The final portion of the study protocol was a semi-structured conversational interview with participants. Participants were instructed to stop sharing their screen. The interviewer proceeded to have a discussion with the participants about their use of EVM in the think-aloud sessions and the design of software incorporating model checking functionality into VA tools. The interviewer followed an interview guide during this portion of the study protocol, addressing each of the following topics in turn by asking the listed questions verbatim:

Topic 1: Utility of model check visualizations.

* In what ways (if any) did you use model checks to help you think about the dataset?
* What specific visual cues on the resulting chart (if any) were interesting or helpful to you?

Topic 2: Generative thinking.

* Did you find yourself thinking about data generating process?
* What kinds of assumptions (if any) did you make about the dataset?
* Did using model checks make these assumptions more salient or concrete?

Topic 3: Expressing models.

* Did you have any difficulty using the model bar to express and check provisional interpretations of data? What if anything made it hard to use?
* What do you think would make this kind of function easier to use?

These questions served as a jumping off point for discussion. The interviewer clarified terminology like “model checks” or “assumptions” as needed. After asking each question, the interviewer took notes on the participant’s response and asked clarifying follow-up questions. For example, when possible, the interviewer brought up concrete examples of instances during the think-aloud sessions where the participant exhibited some use case for the tool or seemed to struggle with specifying or interpreting a model. Revisiting these concrete examples made these debriefing interviews more informative compared to if we had asked merely for participants to speculate about software design in the abstract. In response to follow-up questions, some participants introduced new examples they had previously encountered in their workplace, and we discussed those examples. This semi-structured conversational procedure gave us sufficient flexibility to learn what distinct perspectives each participant had to share with us while also ensuring that all interviews followed the same format and covered the same topics.

(This study protocol was deemed exempt by the IRB at Northwestern University.)