Chapter 9
Model Evaluation

Abstract Model evaluation is examining the appropriateness of a model for a given situation or data. The Focal KSAs in model evaluation tasks are the capabilities to determine whether, how well, or in what aspects, a model is appropriate for a given situation. Potential Observations may include whether students identify cues of model misfit; whether particular areas, patterns, or unaccounted-for features of the situation are identified; and whether hypotheses for the model-data discrepancy can be proposed. The Model Evaluation *design pattern* is tied closely with the Model Use and the Model Revision *design patterns*.

Model evaluation is examining the appropriateness of a model for a situation or data. This may be as straightforward as answering the binary question of whether or not the model fits the data, it may require an explanation, or it may involve an investigation of how well or in what respects the model fits and fails. While tasks can be devised that focus primarily on model evaluation, this aspect of model-based reasoning is intimately connected with several other aspects of model-based reasoning. Model evaluation is tied inextricably with model use. In order to evaluate a model, students must be able to reason through the model to project its facsimile of the salient features of the situation, whether qualitative or quantitative, because comparing these projections with the actual situation is the basis of model evaluation. While it may be hard to separate model use and model evaluation (and often unnecessary), tasks can be designed to focus on model evaluation for more targeted instruction and assessment.

9.1 Rationale, Focal KSAs, and Characteristic Task Features

In any type of model-based reasoning, students need to be able to connect the real-world situation and the model (the arrow in the lower left corner of Fig. 2.1). Without model evaluation, students have no justification for why one model may be

R.J. Mislevy et al., *Assessing Model-Based Reasoning using Evidence-Centered Design*, SpringerBriefs in Statistics, DOI 10.1007/978-3-319-52246-3_9

better than another, and therefore may not be able to determine an appropriate model. In real-world situations where the model is not provided, students will have difficulty addressing the problem if they cannot evaluate, as well as propose, candidate models.

There are three basic ways to elicit evidence about model evaluation in tasks: a model is provided and the student must address its appropriateness; multiple candidates are provided or suggested and the student must determine their suitability; and a model is not given and the student must formulate a model. The first two focus attention on model evaluation specifically, while the third integrates model evaluation into the flow of inquiry. In an investigation task, a rubric can include assessing model evaluation as it occurs in students' ongoing procedures or in their final presentations.

Model evaluation is often prerequisite to model revision and model elaboration; it is necessary to determine how and how well a model fits a situation before one can improve it. The quantum revolution was motivated in part by Newtonian and field mechanics' failure to account for the photoelectric effect and "block box" radiation. Tasks designed to provide evidence about model evaluation can be extended by model revision or elaboration.

A classic example of model evaluation in statistics is multiple regression. A variety of model-checking tools are used to examine how well the model fits and the structure of the relationship between the independent and dependent variables (e.g., Belsley, Kuh, & Welch, 1980). Mosteller and Tukey (1977) show how to use them in inquiry cycles. A simple regression model posits a linear relationship between the two variables, or $y = ax + b$. An analyst may hypothesize that age and strength are related, such that as people grow older they get stronger. To evaluate this linear model, would study the fit of the regression model to data on subjects' ages and strength. One method is to test the theory graphically (note the articulation needed between an equation and graphical representations). Does the pattern in the data points look like what one would expect under the theorized relationship? For age and strength, the researcher may find that the graph looks more curvilinear, or is linear only within ranges, thus moving toward to model elaboration or model revision. Figure 9.1 shows situations where the linear regression model appears suitable, there is a curvilinear relationship that it cannot capture, and an outlier that renders the regression line misleading.

Baxter, Elder, and Glaser's Mystery Boxes (1996) combines model evaluation, model use, and model revision is. Students are presented six different boxes with some combination of elements among a light bulb, wire, and batteries, and they must perform tests to determine what is inside each box (Fig. 9.2).

The students have been studying a model for simple circuits with these components. In this hands-on task, they must use their understanding of this model to determine what sub-model fits each of the boxes. They must determine which tests (connecting the terminals of the mystery box with just a wire, with a battery, with a light bulb, and so on) are appropriate in narrowing down the choices for the submodel. Interpreting the results of a test requires reasoning through each provisional model to predict what would be observed if it were the true configuration (model use), then determining whether the observed result is consistent with the

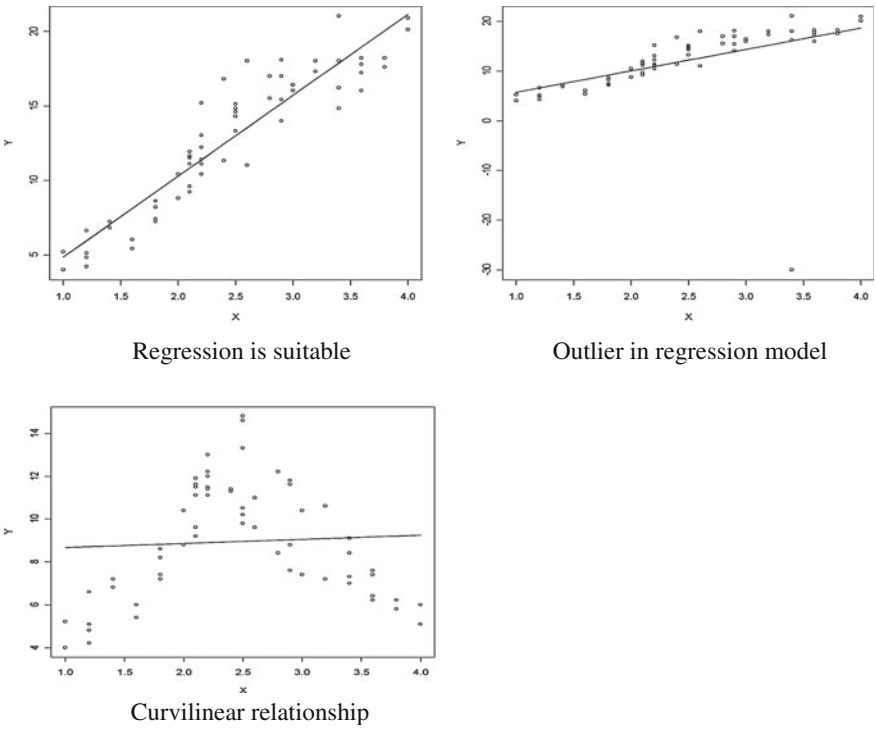

Fig. 9.1 Examples of a simple regression model with three data situations

prediction. Comparing this prediction and what actually happens is model evaluation. Generally a single test is not sufficient to determine conclusively which configuration is inside a box, so the results of multiple tests must be synthesized to evaluate each possibility. This feature of the task leads to some Potential Observations concerning evaluation strategies, which will be discussed in Sect. 9.4.

The Focal KSAs in model evaluation tasks are the capabilities to determine whether, how well, or in what aspects, a model is appropriate for a given situation, be it a real-world scenario or already-synthesized data. This can include identifying relevant features of the data and the model(s) under investigation, and evaluating the degree and nature of the correspondence between them. In the examples, students must be able to examine the data given (the regression data set or the physical mystery boxes) and determine model fit and suitability.

Characteristic Features of tasks designed to assess model evaluation include a target situation and one or more models. The models should be able to be examined in light of the situation and the data. In the regression example, the situation is predicting an outcome variable, and the model is the statistical relationship between the variables. In Mystery Boxes, the situation is determining what circuit in a given box produces the observations the tests produce. The family of models at issue is the set of completed circuits that can be formed from the configurations of elements within the boxes and elements that connect the terminals.

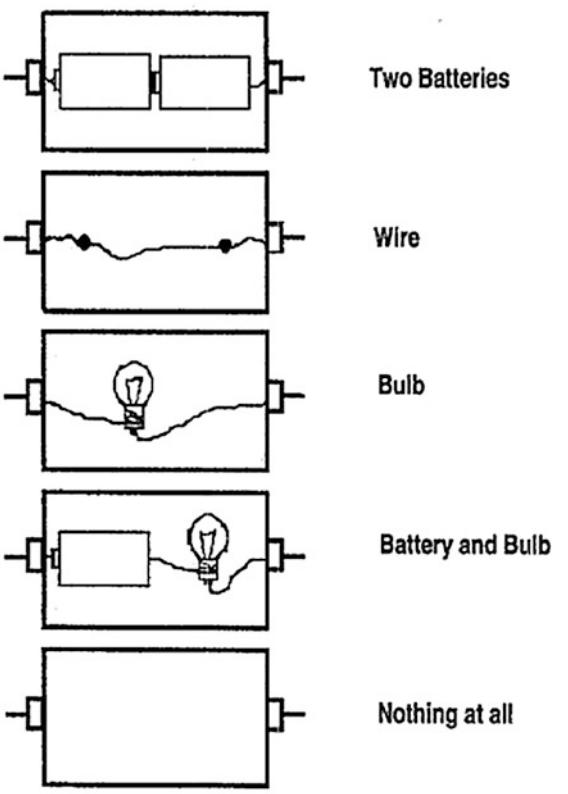

Fig. 9.2 Mystery box task (Baxter et al., 1996)

Box Robotics-4. Model Evaluation

Model Evaluation is central to the robotics task. As in many engineering problems, theory and experience guide the design of an artifact that will produce desired results under given constraints (Simon, 1996), but solutions may require repeated trials and successive approximations. In this task, Model Evaluation requires analyzing, critiquing, and diagnosing the behavior of a simulated or physical rover in each trial. That behavior arises from the configuration being tested in that trial. It is the basis for revising the model

(discussed in the following section). To know how to revise the model (simulated or physical), one must understand how the observed behavior departs from the desired behavior, and reason through the model to determine what characteristics of the artifact produced less-than-optimal results. In this task, this aspect of Model Use is thus embedded in Model Evaluation in every testing cycle. In the simulation phase of the investigation, the student can view the rover's behavior and additionally ask to view generated graphs as in Fig. R4. Is the rover traveling up the ramp? At what rate of speed? Does it stop at some point along the way? Are its wheels just spinning?

The particular Focal KSA in evaluating this engineering model is characterizing the artifact's behavior in the criterion situation, with particular attention to its correspondence to the desired behavior. The Characteristic Feature in each testing cycle is behavior through the model in comparison with the desired behavior. This is so in both the simulation or physical phases. It is the necessary feature of a situation to evoke Model Evaluation. Note that this aspect of model-based reasoning is not isolated as an encapsulated task. Rather, each instance is marked by the student working herself into this situation—perhaps without even recognizing it, and thus failing to carry out the process. Further, because it is a part of each testing cycle, one student might work herself into only one such situation, while another works himself into five of them—all evidence-bearing opportunities for assessing the student with respect to this aspect of model-based reasoning.

Again a critical Variable Task Feature is whether the modeling is carried out in the simulation space or the physical space. (Another potential value of this variable task feature would be to carry out the initial design work with paper and pencil renderings and approximations of behavior through equations.) Both phases entail the Additional KSA of domain knowledge with respect to the electric circuit, motor functioning, and gearing ratios and their implications. However the simulation/physical design variable brings with it a cluster of other Additional KSA demands, associated Variable Task Features, and Work Products. The Additional KSAs concern proficiencies for working in the appropriate space, either simulation tools and affordances or the capabilities to run, observe, and record behavior of a physical rover. (It might be mentioned that one could attach an accelerometer to the physical rover, and obtain more information—and at the same time engage corresponding Additional KSAs.)

Another Variable Task Feature that differs across phases is the amount of scaffolding. Built into the simulation environment is a tool call the Learning Companion—a kind of coach that supports Model Evaluation and the upcoming Model Revision aspect and the overarching Model-Based Inquiry activities. As noted previously, the log file of a student's activity in the simulation space is captured, and is used to keep track of a student's rover design in each trial and its behavior. It counts the attempts, and offers feedback that is likely to be helpful. Figure R5 shows its logic.

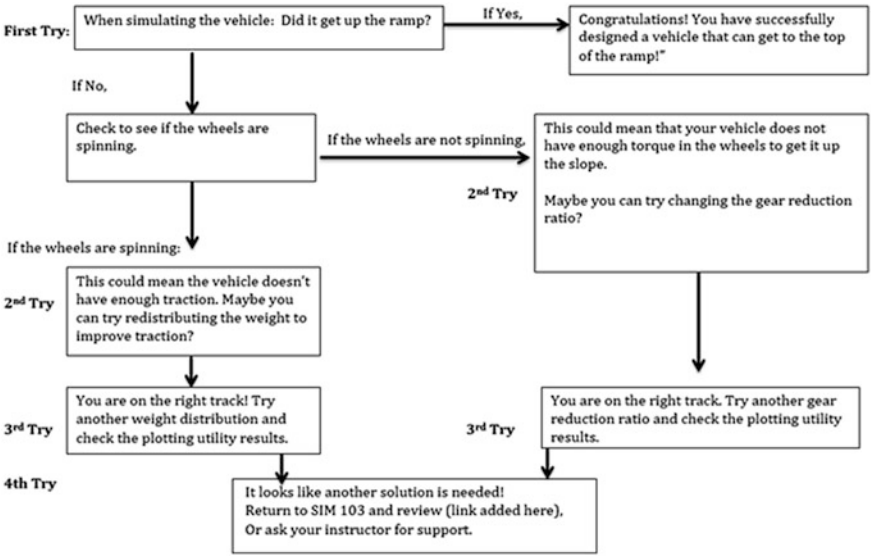

Fig. R5 Learning companion feedback for vehicle trying to climb a ramp

The physical phase of the task does not engage the Learning Companion. The intent is that the support it affords in the simulation phase will have provided the student with a schema of kinds of rover behaviors to look at and what they mean (and later, what to do next).

9.2 Additional KSAs

In addition to the Focal KSAs described above, assessments of model evaluation can require different levels of domain knowledge and familiarity with the type of task or model being evaluated. With regard to domain knowledge, being able to evaluate the fit of the model depends on being able to identify mismatches between a model and a situation. The more subtle the mismatch and the more it depends on the particulars of the model, the more critical the Additional KSA of domain knowledge becomes; domain knowledge sets expectations about what features are relevant and what relevant patterns should look like. Thus an assessment meant to focus on the process of model evaluation per se would use familiar models and situations. An assessment meant to address model revision jointly with knowledge of particular models can validly have a high demand for the substantive aspects of those models.

Additional KSAs also include familiarity with the methods used to evaluate the model, and the standards and expectations in the field. As noted below regarding Variable Task Features, a designer can use scaffolding to reduce the demand on Additional KSAs consisting of background knowledge, planning, and evaluation methods such as graphical and statistical tools in the regression task. As always there are tradeoffs: requiring fit indices to be calculated by hand increase demands for computational procedures, but requiring the use of a particular computer program instead brings in Additional KSAs for using the program.

In the Mystery Boxes task, students' knowledge of circuits materially affects the difficulty of the task. The students in Baxter's study had just completed a unit on circuits, so the evidentiary focus of the task for them was in planning and carrying out the testing procedures to infer what was the boxes. Students who are not familiar with circuits might not be able to reason through possible combinations of elements (model use) to carry out model evaluation (although the tasks could be instructional activities to help them learn about circuits). Students also need some knowledge to connect the components. These Additional KSAs concerning procedures would be circumvented in a simulation version of the task. Further, students could be scaffolded in steps of evaluating the boxes in order to reduce demands on planning and organizational capabilities. Baxter et al. chose not to, because evaluating how the students planned and explained their procedures was central to their research.

9.3 Variable Task Features

Model evaluation tasks can vary as to the type and complexity of the model(s) to be examined. Model evaluation can be prompted or unprompted (implicit) in a given task. That is, tasks can present models to students and explicitly direct the student to examine them, alternatively, students have to determine the models themselves, and indeed whether or not to evaluate fit. Whether the model at issue fits, and if not, the degree and nature of misfit, can also be varied. The type and complexity of the model evaluation methods may differ. These choices can be used to increase or decrease demands for particular aspects of the Focal KSAs and Additional KSAs. Different choices can provide more or less information, often trading off against convenience and economy of scoring procedures. More open-ended tasks take longer for students to complete and present more challenges for scoring, but provide more evidence about students' reasoning, and incorporate model evaluation into the inquiry process.

In regression tasks, for example, the number of predictors can be varied. The students can also just be asked to use graphical displays to explain why they believe the model fits or does not fit, or to use statistical methods or graphical methods to justify their conclusions.

In Mystery Boxes, the medium of the task could vary: physical boxes, an interactive computer simulations, or static paper-and-pencil representations. The last is simplest and easiest to score, but it places a greater demand on model use for

projecting the results of different configurations, as the task environment itself no longer provides feedback. The number and contents of the boxes could be modified, as well as the information students have about what they might contain. All of these modifications can affect task difficulty and raise or lower demands on different aspects of knowledge, both focal and additional.

Mystery Boxes illustrates design choices about the kind and amount of scaffolding to provide. As mentioned, Baxter et al. did not scaffold the process so they could obtain evidence about students' planning and self-monitoring. A different scaffold would be a chart of the results of tests when applied to different configurations. It removes most of the demand for reasoning through the circuit model and shifts the focus to model-evaluation procedures. Whether to do so depends on the intended examinees and the purpose of the task.

9.4 Potential Work Products and Potential Observations

The simplest Work Product for a model evaluation task is the indication (in whatever format specified) of whether or not the model fits or which model fits. It is also least informative. The next more informative option is having the student provide qualitative and/or quantitative indications of degree and nature of fit and misfit. Statistical tests, graphs, or other representational forms for model evaluation can be evoked as Work Products. These "final product" Work Products can be accompanied by an explanation (verbal or written) of why and how the student reached the conclusion. This can include verbal or written explanations of the hypotheses formulated (regarding fit) and the methods used to test them, including the output from model-fitting tools. Written, verbal, or computer-tracked traces of the actions a student performed can also be captured as Work Products.

Compared with simple choice Work Products, explanations are particularly useful in determining if a student understands the situations and models well enough to evaluate them critically. The formality of an assessment may dictate the format required as well as the depth expected. Solution traces can take various forms, such as computer logs of actions students take through a computer interface in a simulation investigation, a video recording of actual performance, or a written trace by the student of the steps in their evaluation. All of these examples provide more evidence about the efficiency of the student's model evaluation procedures than a final solution. Baxter et al. (1996) found the last of these particularly useful for evaluating the rationale behind students' decisions.

Potential Observations in model evaluation tasks can thus address the comprehensiveness and the appropriateness of the hypothesis generated through the model for evaluation, the appropriateness of the evaluation method(s) used to assess model fit, the efficiency and the adequacy of procedures the student selects, and the correctness and thoroughness of the evaluation. This can include whether students identify cues of model misfit; whether they identify particular areas, patterns, or unaccounted-for features of the modeled situation; and whether they propose

hypotheses for model-situation discrepancies. All these Potential Observations provide evidence about model evaluation in context, but all also require some degree of understanding of the substance of the situation, as Additional KSAs.

More complex Work Products provide the opportunity to explore these qualities more deeply. Simpler Work Products—such as selection of a best-fitting model—provide less information but, on the other hand, allow the aspect of proficiency to be targeted more precisely. The quality of the explanations given, as well at the quality of the determination of how an ill-fitting model might affect inferences resulting from that model, also can be examined. How well a student integrates results from multiple methods of testing can be observed, along with how well a student is able to indicate which aspects of the model and data do not fit.

In the regression example, the Work Products can include the output from a formal model fitting tool and an explanation of the conclusions drawn from observing this output. From the output of graphical and statistical model-fitting tools, an assessor can evaluate the quality of the explanation and the appropriateness of the tools used and how they were applied.

In Mystery Boxes, Baxter et al. gathered as Work Products the students' initial plans, their strategies, and explanations of their solutions, and traces of their activities in the form of written logs and think-aloud protocols. The researchers evaluated the "explanation" Work Products for what students expected if a certain combination of components was inside the box—the aspect of model use that is integral to evaluating a proposed model. From the trace of students' activities, the investigators observed and evaluated how flexible the students were as they monitored their results. The Work Products made it possible to create observations that addressed not only the end results (the students belief about the contents of each box) but also how well the students were able to interpret the results of each of the tests to determine which further tests, if any, were needed.

9.5 Some Connections with Other Design Patterns

The Model Evaluation *design pattern* links closely with model use. It can even be difficult to develop tasks that assess only model evaluation. However, tasks can be designed to emphasize either model use, model evaluation, or both. This may be accomplished by scaffolding whichever aspects of reasoning (if any) are not the intended focus of the task (e.g., providing a table of test results in the Mystery Box tasks).

The Model Evaluation *design pattern* also is associated with model revision and model elaboration because in order to determine if a model needs to be revised or elaborated, some model evaluation usually needs to have been performed. Model revision tasks can be designed to minimize model evaluation by presenting the students with a situation and a model they are told is inadequate in a given way, which they need to revise or elaborate.