Office of Consultation and Research in Medical Education (OCRME)

Educational Research and Evaluation Workbook

A Self-Study Guide for Completing an EDF Proposal Evaluation

Rick Axelson, Ph.D.
OCRME, Program Evaluation Consultant

Susan Lenoch, M.A.
OCRME, Associate Director & Instructional Design Consultant

Last Revised: March 4, 2009
OCRME Educational Research and Evaluation Workbook

Contents

Purpose and Introduction

Step 1: Describing and analyzing your program—the Logic Model
- What is a logic model?
- Value of a logic model
- Creating an initial logic model

Step 2: Refining your study design and evaluation questions
- Did the program work? If not, why not?
  - Examining outcomes and their linkages with program activities
- Was the program implemented as intended? If not, why not?
  - Examining program activities
- How could the program be improved?
  - Examining goals, assumptions, and external factors

Step 3: Gathering evidence to answer your questions
- Measures
- Relationships
- Study Design Considerations--Putting these Elements Together

APPENDICES

Appendix A. CLIPP Case Summary
Appendix B. Sample Method for Developing and Prioritizing Evaluation Questions
Appendix C. The Kirkpatrick Model of Program Outcomes
Appendix D. Reviewing Quantitative Educational Research or Evaluation Studies
Purpose and Introduction

This manual provides a simple, straightforward guide for designing a research or evaluation study. It is intended for those who wish to study an educational program\(^1\) or intervention but have little formal training in educational research or evaluation methodology. There are many excellent textbooks and manuals on these subjects but their level of technical detail often hinders the newcomer, intent on getting their project underway.

The following chapters provide stepwise guidance to help you systematically and thoughtfully plan a simple study\(^2\) of high quality. We hope to provide “just enough” explanation of the theories and technical processes underlying study design to help you see the logic of it all and to assist you with identifying issues where you might benefit from technical consulting or assistance.

This manual will also assist you if you are applying for a Carver College of Medicine Educational Development Fund (EDF) grant. These proposals must include an evaluation plan.

\(^1\) For our purposes, a program is any purposeful and organized effort to intervene in order to solve a problem, accomplish a goal, or provide a service. As you can imagine, medical education programs can be very simple, consisting of a single activity or intervention, or very complex.

\(^2\) We don’t make a distinction between research and program evaluation in this manual. Although evaluation studies tend to focus on program improvement or assessing outcomes, they can, like research studies, focus on understanding why or how a program works. Further, there are no methods or designs that are distinctly within the sphere of “research” or “evaluation”—rather methods and designs can be classified by their usefulness for the particular questions posed in the study.
Step 1
Describing and Analyzing Your Program: The Logic Model

A thoughtful analysis of what your program is, what stage of development it is in, and how you think it should work, is a good starting point for designing your evaluation. This analysis leads to a clearer conception of the program and the information that is needed to enhance or evaluate its performance. One tool often used to facilitate program analysis and description is the logic model.

What is a Logic Model?

A logic model is a map or diagram showing your program components and your underlying rationale about how and why you think the program will work. It begins with a description of the program but it is more than that because it attempts to make explicit your assumptions and your hypotheses about how program components interrelate and result in expected program outcomes. It helps posit connections between program activities and outcomes which can form the basis of your study questions.

See the illustration on the next page of a simple logic model for a “program” to buy a car.

The components of a logic model are as follows.

Goals: Statements about the purpose of your program or intervention. What problem are you addressing with this program? What are the goals and objectives? (For educational programs, these could be learning goals: what your participants will know, or be able to do upon completion of training.)

Assumptions: Untested beliefs about the program, the participants, the conditions you will encounter, and how you think the program will work. Assumptions could also include theory or hypotheses derived from previous research that apply to this study.

Inputs: The resources available for your program including human (e.g., participants in the program, staff who will implement it), financial, technical, physical and organizational. The emphasis here is on identifying key factors that are essential in implementing your planned activities.

Activities or interventions: These are the main, as well as adjunct, activities, actions or interventions of your program for which you will need the resources. If you have multiple target audiences, you can group the activities according to whom they will be directed.

External Factors: These are factors outside your program that could affect your program. Often you can’t control these factors, but you may be able to measure their effect.
Outcomes: “Outcomes” and “impacts” have distinct meanings for some evaluators but they both refer to the effects that your program may have on participants and the larger organization or society. Among evaluation models, outcomes most commonly refer to the products of the program (sometimes called “outputs”) and to changes in participants or subjects that are more directly and often more immediately associated with the program. In an educational setting, outcomes refer to changes in participants’ learning and behavior. Impacts are longer-range or longer-term effects of the program on participants, or refer to effects that participants have on others, or on the larger organization, environment or society. Impacts typically take time to become evident and are subject to greater numbers of factors that are beyond your control as a program planner or researcher. You can think of outcomes and impacts as similar to a ripple in a pond. The outcome is the initial splash the pebble makes, while its’ full impact could reach the further side of the pond.

For example, a hospital in-training program on proper handwashing may have an immediate effect on program participant’s observable handwashing behavior (learning outcome) and on their routine behavior on the job (transfer of learned behavior to practice). Eventually this improved behavior may affect the rates of hospital-transmitted illness but this impact may take time to develop since it depends on having a critical number of staff performing proper handwashing. A further impact on the hospital organization might be lower costs, improved reputation, etc.

### Illustrative Logic Model: Buying a Car

**Program Description:** Family will purchase a 2nd family car.

**Goal(s):** Buy an affordable car that meets our transportation needs; Send car off to college with our child

**Assumption(s):**
- We need a 2nd car
- You can’t get everything you want in a car (at an affordable price)
- Teenage drivers increase the cost of owning a car
- Now is a good time of year to get a deal on a car
- Following a systematic decision-making process will result in a better purchase than one driven by emotion.

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Activities/Intervention</th>
<th>Outcomes: Short-term</th>
<th>Medium-term</th>
<th>Long-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ to put down</td>
<td>Generate and prioritize car features we need</td>
<td>Buy a car that meets our transportation needs</td>
<td>Pay off car in 4 years</td>
<td>Car lasts 9 or more years</td>
</tr>
<tr>
<td>$ for trade in of old car</td>
<td>Web-shop possible cars</td>
<td>Affordable car payments</td>
<td>Send car to college with child</td>
<td></td>
</tr>
<tr>
<td>Credit available</td>
<td>Create short list of cars that meet needs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family members including 14-year old</td>
<td>Test drive from short list</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 current cars (age, features, reliability, cost)</td>
<td>Select top 2 choices</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resources to research car makes and models</td>
<td>Negotiate best deal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insurance/licensing costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Potential factors that could affect the success of your program (for good or bad):
Ability to stick to the plan
Negotiation skills
Car is/is not a “lemon”
Child is/is not a good drive
Car accident(s)

In the “buying a car” logic model the inputs include all the resources, constraints, and considerations the family takes into account in making a decision about a car. The activities describe the process they will use to try to make an objective decision, first by prioritizing what features they need in the car, reviewing options available and then narrowing choices before test-driving. The outcomes include a process goal (following the plan) and both immediate and longer-term desired outcomes. They also identified things that could thwart their plans and in doing so, may be able to take steps to alleviate them. Listing assumptions helps them take a step back and reflect on the proposed program and goals to see if there are other paths to be considered. In this case, family members see that they have not considered transportation options other than a second car. Explicitly stating goals and assumptions helps assure that all family members are working together and considering all available means to meet their shared goals.

Value of a Logic Model

The value of constructing a logic model prior to designing your study is that the process helps you clarify what you are trying to do with the program and identify any gaps in your thinking or planning. From the model, you can develop explicit statements or questions in the form of testable causal statements. Multiple “if/then” hypotheses could be proposed for a given logic model by “reading” across the diagram, e.g., “If I have the resources I need I can accomplish the activities I propose.” “If I implement the activities as proposed participants will benefit from new knowledge and skills.”

A logic model is useful in designing a study but also in designing the program itself and communicating with others about the program. It helps you focus on these important questions:

- Are the goals of your project realistic?
- Are your assumptions and expectations valid?
- Have you identified potential unintended outcomes?
- Are the activities of your program logical in relation to the goals?
- Have you identified sufficient resources?
- Does everyone involved agree with this logic model? (If some of your stakeholders are at odds with your model, you may have significant obstacles.)

Creating an initial logic model (goals, inputs, activities, outcomes)

In this workbook, we will create a logic model for an educational program in two steps. Our initial logic model, below, will focus on what we plan to do to achieve our program
goals. In developing this version of the model, we will use only four model components (goals, inputs, activities, and outcomes). In the next section, we will refine our logic model by adding assumptions and external factors to articulate why we expect the model to work.

For our initial model, it is reasonable to start with a description of the program and continue in the order defined by the chronology of how things will occur (see worksheet on the following page). However, you can start with any of the above components. For example if you are implementing a new training program for the purpose of achieving specific outcomes, you will want to start by defining those outcomes so that everyone is clear on what it is you are trying to accomplish. If your interest is experimenting with a new teaching strategy you may want to begin by thoroughly describing the strategy under “activities.”

**CLIPP Case Study**

To illustrate the evaluation design process discussed in the workbook, we will use an EDF project completed by Dr. Michael Tansey and Dr. Jerold Woodhead from the UI Pediatrics Department. Below is the abstract from their EDF proposal. (See Appendix A for a more detailed project description.)

We propose to test the effectiveness of the Computer-assisted Learning in Pediatrics Project (CLIPP) during the Pediatric Clerkship for third year medical students. CLIPP is a web-based program that covers the General Pediatric Clerkship Curriculum objectives. At the beginning of the clerkship we will assess baseline knowledge by administering a pre-test, including questions covered by the CLIPP cases. We will then ask students to complete at least six CLIPP cases during the clerkship. We will monitor each student’s use of CLIPP, noting the specific cases covered, the time spent on each case, and whether the case was completed or partially done. At the end of the clerkship, each student will take a final exam that includes questions covering the areas of knowledge of the CLIPP cases. We hypothesize that student use of CLIPP will be positively correlated with a gain in score when comparing pre-test to final exam results.

Some other key points about the program from their EDF proposal include:

- CLIPP is designed to supplement traditional clerkship and patient care activities. It provides access to peer-reviewed learning materials.
- CLIPP has 31 interactive virtual patient cases that are designed to cover all of the learning objectives of a standard core curriculum in pediatrics.
- The cases are interactive, incorporate multimedia extensively, and frequently require the student to make decisions about diagnoses and clinical management with support from an expert – the case author.
- CLIPP models and teaches diagnostic reasoning. At key points in the case the student is required to develop a differential diagnosis based on the case findings.
- CLIPP provides a bank of test questions that are based on brief clinical vignettes. The test questions cover the cases that will be used at the University of Iowa.
Below is an initial logic model showing the four basic program components.

**CLIPP Initial Logic Model**

**Goal(s):** to effectively incorporate web-based cases as an adjunct learning activity in the pediatrics clerkship

<table>
<thead>
<tr>
<th>The Program: Inputs</th>
<th>Activities/Intervention</th>
<th>Outcomes: Short-term</th>
<th>Medium-term</th>
<th>Long-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>140 third-year medical students (at different times of the year)</td>
<td>Pre-test of students’ topic knowledge based on clinical vignettes</td>
<td>Students’ knowledge of topics and ability to apply it to clinical cases will increase with the use of CLIPP cases</td>
<td>Students will become better self-directed learners through their experience with the CLIPP exams and cases</td>
<td>Students will provide better clinical services in pediatrics due to improved learning and retention of knowledge</td>
</tr>
<tr>
<td>31 CLIPP cases previously peer and student-reviewed</td>
<td>Students select 6 or more CLIPP cases that they study during their clerkship</td>
<td>Students will enjoy the self-directed learning process of CLIPP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database with clinical vignette questions for each CLIPP case</td>
<td>Post-test of students’ topic knowledge based on the clinical vignettes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current clerkship learning activities – readings, didactics, patient care</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Availability of computers/access to web-based cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Now you try.**

Create an initial logic model of your own program using the worksheet below. You may find that creating the logic model reveals some gaps or inconsistencies in the program. That’s okay—one value of a logic model is revealing gaps so they can be fixed before the project begins.
**Step 1. Initial Logic Model Worksheet**

Goal(s):

<table>
<thead>
<tr>
<th>The Program: Inputs</th>
<th>Activities/ Intervention</th>
<th>Outcomes: Short-term</th>
<th>Medium-term</th>
<th>Long-term</th>
</tr>
</thead>
</table>

Now that you have described what your program looks like and what it intends to accomplish, you are ready to take a closer look at issues related to the effective functioning of your program, issues that are revealed by examining your program components more deeply. Step 2 will lead you through this process.
Step 2
Refining Your Study Design and Evaluation Questions

For your EDF project, as is the case with most program evaluations, three of the most critical issues that you will likely need to address are:

1) Did the program work? If not, why not?
2) Was the program implemented as intended? If not, why not?
3) How could the program be improved; that is, how could it become more effective and/or efficient?

We will use these three questions to help us refine our program design and clarify our evaluation questions. Our method for doing so builds upon the program logic model developed in Step 1.

Addressing Question 1: Did the program work? If not, why not?

To address this question, the “Outcomes” columns of the Program Logic Model provide a good starting point. An effective program is one that produces the desired outcomes. Although determining whether the desired outcomes occurred is an essential component of the evaluation, it is equally important for evaluation design purposes to have a clear understanding of how the program is expected to produce the outcomes.

As an example, we will elaborate how the CLIPP program activities (independent variables) are expected to produce the desired outcomes (dependent variables). How will this help us design the evaluation? It helps in a few ways. First, a clearer understanding of how the program activities affect outcomes will enable us to improve the program and adapt it to our learners and their educational needs. Secondly, an elaborated causal process lends itself to a more convincing argument about the program’s activities as the cause of the observed outcomes. Demonstrating a sequential path of events leading from initial cause through intervening mechanisms to observed effects is more informative than simply showing a correlation between the program activities and the observed outcomes. Thirdly, if the program does not work as intended, we will have a good foundation for explaining what went wrong. Our first task in this section will be to clarify the linkages between program activities and outcomes. These might be thought of as intervening mechanisms or mediating variables that account for the relationship between the independent variables (activities) and the dependent variables (outcomes).

3 It is also important for program improvement and generalizability purposes to know more detail about instances where the program worked. Why did it work? For whom did it work? Under what circumstances did it work? These issues will be addressed in our discussion of program improvement (Question 3).

4 The present discussion will be limited to the three typical issue areas noted above. For a more comprehensive approach, see the method for surfacing all of the relevant information needs for your evaluation study outlined in Appendix B.
From the CLIPP case, we start by examining one of the short-term outcomes from the initial logic model.

- Students’ knowledge of topics will increase with the use of CLIPP cases

In their EDF proposal, Tansey and Woodhead cited evidence of CLIPP’s effectiveness from six schools where it had been used. They noted that most students found CLIPP to be more effective than reading or other traditional “didactic” methods. Further, student descriptions of the positive aspects of CLIPP suggest how it may be effective. Students reported that it provided solid general pediatric knowledge and engaged them in active learning. They thought it improved their ability to evaluate (generate a differential diagnosis) and manage common problems.

From this and other information about the nature of the CLIPP cases we can derive some hypotheses about how CLIPP cases would likely facilitate student learning at our school. It would seem that CLIPP cases might be effective because they provide

- High quality content. CLIPP cases appear to provide logical, targeted, and/or coherent presentations of the relevant information that in turn improve students’ comprehension and retention of the material.
- More active learning. The format appears to improve students’ learning by asking them to perform tasks in which they must use the content they are learning in application to clinical vignettes.
- Focused instruction on clinical reasoning. Active learning tasks include developing a differential diagnosis based on case findings and justifying their diagnosis by showing whether findings argue for or against their differential diagnosis.

Depending upon which features of the CLIPP material aid improved learning, we would expect students to display different patterns of CLIPP case usage and report alternative reactions to their study experiences.

For present purposes, however, our focus will be on specifying possible mediating variables and mechanisms. In thinking about mechanisms, it is often helpful to consider what you are hoping that students experience from the program (e.g., specific types of thoughts, feelings, and/or behaviors) and how those experiences will lead students to produce the desired outcomes. Next, consider what elements or aspects of the program are intended to produce these responses in students.

In the following table, we illustrate this approach with the CLIPP program. Our initial logic model has been expanded by adding “mechanisms” showing linkages between interventions and outcomes. We have split the mechanisms into two parts, paralleling the above discussion about student responses and the program elements that are intended to produce them. We have posited that student use of CLIPP cases would improve learning through three possible intervening mechanisms.
## Expanded CLIPP Logic Model with Intervening Mechanisms

<table>
<thead>
<tr>
<th>Interventions/Activities</th>
<th>Mechanisms</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elements of CLIPP activity that affect student response</td>
<td>Student Responses that lead to the outcome</td>
<td>Short-term</td>
</tr>
</tbody>
</table>

- **Students select 6 or more CLIPP cases that they study during their clerkship**
  - High quality content
  - More active learning
  - Focused instruction on clinical reasoning
  - **Students develop better comprehension of material**
  - **Students become better at applying knowledge to vignettes**
  - **Students’ enhanced clinical reasoning skills enable them to better assess evidence**
  - **Students will have significantly better performance gains on post-tests when they study with CLIPP cases**

---

### Now you try.

Using your Program Logic model, elaborate the linkages between a program activity and an outcome. Identify intervening mechanisms that will enable the program activities to produce the desired outcomes. Write your proposed mechanisms under each of the items in the “mechanisms” column as shown in the CLIPP example above.

<table>
<thead>
<tr>
<th>Interventions/Activities</th>
<th>Mechanisms</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elements of activity that affect student response</td>
<td>Student Responses that lead to the outcome</td>
<td>Short-term</td>
</tr>
</tbody>
</table>
The value of this exercise becomes especially evident when the program does not work as intended. In that event, we will have two sources to check for possible explanations.

1) Did the program activities activate the posited intervening mechanisms?
2) Were the program activities properly implemented? (This is the issue we examine in the next section.)

Note, in the foregoing example, we considered only one of the program’s short-term program outcomes. For a complete program evaluation, you would want to consider the full range and extent of possible outcomes in the short- (immediate and direct), medium-, and long-term. Consider both the outcomes you hope will happen but also unintended outcomes.

Kirkpatrick has developed a model to examine levels of program effectiveness that is useful for educational and training programs. He suggests that program effectiveness can be measured at four levels: 1) participants’ reactions to the program; 2) their learning from the program; 3) their behavior in practice (e.g., on the job) as a result of learning; and 4) the longer-range effects of the program on participants, on those who interact with them (e.g., their patients), or on the organization as a whole, as a result of participants’ changes in behavior. For more about this model, see the supplemental chapter in Appendix C.

**Addressing Question 2: Was the program implemented as intended? If not, why not?**

We now turn our attention to the “Interventions/Activities” column of the logic model. Let’s use the activity/intervention listed in the above CLIPP Logic Model to focus our discussion. That is,

- Students select and utilize a minimum of 6 CLIPP cases to augment clerkship studies with self-directed learning.

To address the issue of whether the CLIPP program was properly implemented with respect to this activity, we will not only need to determine whether the above events occurred, but whether they did so in a manner needed to activate the intervening mechanisms. In our present case, one could ask about students’ utilization of the CLIPP cases. Did students put sufficient time and effort into self-directed study with the CLIPP cases to affect their knowledge gains?

As you can see, questions like this may lead you back into the research literature to identify best practices in facilitating CLIPP case use for maximum student benefit and may result in changes in your program design. At the conclusion of this iterative process, you will have a refined set of evaluation questions that will enable you to assess whether the program has been properly implemented. For example, with the CLIPP case, we could decide that students must meet certain time and effort thresholds. The treatment thresholds might take the form of students being required to log a certain minimum amount of study time per CLIPP case section, complete all sections of the case, and correctly answer a specified percentage (e.g., 70%) of each section’s tutorial questions. In
our study design, we could then build in ways of ensuring that these standards are met or develop means for capturing data to assess the time and effort students put into studying the cases.

Another way of thinking about your program activities is as a set of independent variables that you believe will produce the desired program outcomes. In asking whether the program was implemented as intended, you are in effect asking whether the independent variables were activated. An obvious, but often overlooked, cause for program failure is that the independent variables were not sufficiently activated to produce the desired outcomes. Fleshing out questions about your program activities will help ensure that data is available to investigate these issues.

To summarize, we have now added some supplementary questions to the CLIPP case that will help us to assess the fidelity of the implementation to the intended program design.

- Students select and utilize a minimum of 6 CLIPP cases to augment clerkship studies with self-directed learning.
  - Did students select and work on a sufficient number of CLIPP cases (i.e., 6 or more) during their clerkship?
  - Did students put sufficient time and effort into self-directed study with the CLIPP cases to affect their knowledge gains?

If the program was not implemented as intended, we will have two general sources of information to draw explanations from. First, in the column of the logic model labeled “Inputs” we have a list of the needed resources to support the program. Failure to obtain one or more of these resources may account for implementation problems. Secondly, areas where program activities were not implemented as intended could be identified through the analysis of data related to the above program activities questions.

Here, we considered only one of the program’s activities. For a complete program evaluation, you would want to examine all of the program activities in developing your supplementary questions.

- Now you try.

Using the intervention activities in your Program Logic Model, think about the types of questions and data you might need to track whether the activities are being implemented as intended for your program. Write your supplementary questions under each of the program activities.
Addressing Question 3: How could the program be improved; that is, how could it become more effective and/or efficient?

The remaining components of the logic model (goals, assumptions, external factors) are good starting points for addressing this question. Let’s begin by reexamining your program’s goals. Do they fully capture the range of benefits the program is expected to yield? After thinking about the program activities and outcomes, as we have in the last two sections, you might change your mind about the reasonableness of the program’s scope and expected outcomes. Or, conversely, you might find that there are additional learning opportunities that weren’t apparent earlier. For example, with the CLIPP cases, we might consider adding feedback sessions with faculty or residents to reinforce the roles of self-assessment and reflection in students’ self-directed learning processes and thus slightly expand the scope of the project with added curriculum and additional learning outcomes. At the conclusion of this process, you’ll want to make sure that your specified program outcomes accurately and comprehensively reflect your program goals.

Earlier, we elaborated the causal process that is expected to produce the desired outcomes. To check on the completeness of our formulation, we should consider if there are any assumptions about how our program works that need to be addressed. Some examples from the CLIPP case are:

Assumptions:
• Content learned in CLIPP cases is in line with UI CCOM curriculum
• CLIPP provides knowledge appropriate level materials for third-year UI medical students
• Students will put forth sufficient time and effort to learn from the CLIPP cases
• Students will select appropriate CLIPP cases, i.e., ones that enable them to address gaps in their knowledge and skills

The last two items suggest assumptions about students’ abilities to be self-directed learners that we many want to investigate further.

Another aspect to consider in studying program effects is the extent to which other contextual factors inhibit or amplify effectiveness. These could be factors that are known in advance or they could be unforeseen events that impact program delivery. Although we don’t see explicit mention of any “external factors” in the CLIPP EDF proposal, some examples of possible external factors for this case might be the following.

External Factors:
• Students have accessibility to web and the CLIPP cases. Technical problems that hamper connection and web accessibility of materials could decrease the program’s effectiveness.
• Other clerkship activities events that compete for M3’s time could decrease the time and effort they put into CLIPP cases and thus diminish the program’s effectiveness.
Additionally, throughout the program’s implementation and operation, we will want to be attuned to any unanticipated changes in program resources, external events/programs, or student characteristics and faculty personnel that may impact program delivery and effectiveness. Information about such factors is important to include in your analysis of program effectiveness.

Our revised logic model and questions for the CLIPP Program are shown on the following page.

---

**Now you try.**

Review the goals and outcomes in your logic model. Do the goals need to be revised? Do the relationships between goals and outcomes need to be updated?

List the assumptions and external factors in your logic model. What other factors related to context, resources, or program participants could impact the program’s effectiveness? Are there any areas that should be closely monitored for potentially influential, yet unanticipated, events?

**Assumptions:**

**External Factors:**

Based on the above considerations, make any needed additions or changes to your logic model.

---
You now have a refined logic model and set of evaluation issues and questions. You can use the template provided at the end of this section to revise your logic model diagram. As a final check, ask yourself the following questions about your program and model.

- Have you included all the essential goals, program components, and outcomes?
- Are all the outcomes important? ( Seems an obvious question but it guards against bias towards easy-to-measure outcomes over relevance.)
- Have you specified the conditions/questions that will enable you to determine whether the program has been properly implemented?
- Have you elaborated the intervening processes and mechanisms that link program activities and outcomes?
- Have you considered program design assumptions and external factors (anticipated, unanticipated) that could influence the program’s effectiveness?
- Have you taken full advantage of the existing research literature to clarify how, for whom, and under what conditions, your program should work?
**CLIPP Logic Model**

**Goal(s):** to effectively incorporate web-based cases as an adjunct learning activity in the clerkship

**Assumption(s):**
- Students understand how to engage in self-directed learning with the CLIPP cases
- CLIPP provides knowledge appropriate level materials for third-year UI medical students
- Students will put forth sufficient time and effort to learn from the CLIPP cases
- Students will select appropriate CLIPP cases, i.e., ones that enable them to address gaps in their knowledge and skills

<table>
<thead>
<tr>
<th>The Program:</th>
<th>Activities/ Intervention</th>
<th>Mechanisms</th>
<th>Outcomes: Short-term</th>
<th>Medium-term</th>
<th>Long-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inputs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>140 third-year medical students (at different times of the year)</td>
<td>Students select 6 or more CLIPP cases that they study during their clerkship</td>
<td>High quality content</td>
<td>Students will have significantly better performance gains on post-tests when they study with CLIPP cases</td>
<td>Students will become better at self-directed learning through their experience with the CLIPP exams and cases</td>
<td>Students will provide better clinical services in pediatrics due to improved learning and retention of knowledge</td>
</tr>
<tr>
<td>31 CLIPP cases previously peer and student-reviewed</td>
<td>Current clerkship learning activities – readings, didactics, patient care</td>
<td>More active learning</td>
<td>Students become better at applying factual knowledge to vignettes</td>
<td>Students will become more likely to initiate self-directed learning activities to fill knowledge gaps</td>
<td></td>
</tr>
<tr>
<td>Availability of computers/access to web-based cases</td>
<td>Database with clinical vignette questions for each CLIPP case</td>
<td>Focused instruction on clinical reasoning</td>
<td>Students’ enhanced clinical reasoning skills enable them to better assess evidence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Potential external factors that could affect the success of your program (for good or bad):**
- Students have accessibility to web and the CLIPP cases. Technical problems that hamper connection and web accessibility of materials could decrease the program’s effectiveness
- Other clerkship activities events that compete for M3’s time could decrease the time and effort they put into CLIPP cases and thus diminish the program’s effectiveness
## Your Program’s Logic Model

### Goal(s):

### Assumption(s):

<table>
<thead>
<tr>
<th>The Program: Inputs</th>
<th>Activities/ Intervention</th>
<th>Mechanisms</th>
<th>Outcomes:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Program</td>
<td>Student Response</td>
</tr>
</tbody>
</table>

Potential external factors that could affect the success of your program (for good or bad):
**Step 3**

**Gathering Evidence to Answer your Questions**

In the previous sections, we developed a conceptual model of your program and identified some key questions about its performance. Our goal in this step is to identify ways of producing credible answers to your questions. To do so, we will need to specify linkages between your conceptual model and the model’s operation in practice. The primary elements needed for this are: 1) empirical measures of the program elements and 2) hypotheses about the expected relationships between the measures.

**Measures: Linking empirical observations with constructs**

In our logic model, we outlined the constructs that will need to be measured. To translate this conceptual model into an empirical one, we begin by listing the elements of the logic model and the information that would help us see and assess these constructs in the operation of the program. Don’t worry about the accuracy or feasibility of the measures at this point. We’ll refine this list later.

In the tables below, we list possible measures for the CLIPP activities, mechanisms, and external factors discussed in Step 2.

**Activities/Interventions: Independent Variables**

<table>
<thead>
<tr>
<th>Questions/Constructs</th>
<th>Relevant Information/Measures:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students select 6 or more CLIPP cases that they study during their clerkship:</td>
<td>Examine students' login and test records to determine how many cases they completed</td>
</tr>
<tr>
<td>➢ Did students select and work on a sufficient number of CLIPP cases (i.e., 6 or more) during their clerkship?</td>
<td>Number of minutes students were logged onto the web-based applications (time)</td>
</tr>
<tr>
<td>➢ Did students put sufficient time and effort into self-directed study with the CLIPP cases to affect their knowledge gains?</td>
<td>Proportion of tutorial questions answered for each case</td>
</tr>
</tbody>
</table>

---

5 Review of previous research or evaluation studies can be most helpful at this point. Among other things, you will want to note measures that have been used to assess the constructs of interest in your program and any information that previous studies provide about the measures reliability and validity for the populations that were assessed. It is important to bear in mind that the reliability and validity of measures can vary depending upon the population being studied. For example, measures that work well on M1s might not work very well on residents. Appendix F outlines an approach for extracting, synthesizing, and evaluating quantitative educational studies.
Mechanisms: Mediating Variables

<table>
<thead>
<tr>
<th>Questions/Constructs: CLIPP activity that affect student response</th>
<th>Information/Measures: CLIPP activity that affect student response</th>
<th>Questions/Constructs: Student Responses that lead to the outcome</th>
<th>Information/Measures: Student Responses that lead to the outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality content</td>
<td>Professional Pediatric organization’s certifying quality/appropriateness of materials</td>
<td>Students develop better comprehension of material</td>
<td>Students’ satisfaction ratings of the CLIPP case content – clarity and appropriate knowledge level for them</td>
</tr>
<tr>
<td>More active learning</td>
<td>Professional Pediatric organization’s certifying pedagogically valid learning tasks within the cases</td>
<td>Students become better at applying knowledge to vignettes</td>
<td>Students’ satisfaction ratings with the interactivity of format and application exercises</td>
</tr>
<tr>
<td>Students receive instruction and training in clinical reasoning</td>
<td>Student quiz scores on tutorial material covering clinical reasoning</td>
<td>Students’ enhanced clinical reasoning skills enable them to better assess evidence</td>
<td>Analysis of post-test data for evidence of improved understanding</td>
</tr>
</tbody>
</table>

External Factors: Contextual Variables

<table>
<thead>
<tr>
<th>Questions/Constructs</th>
<th>Relevant Information/Measures:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students have accessibility to web and the CLIPP cases. Technical problems that hamper connection and web accessibility of materials could decrease the program’s effectiveness</td>
<td>Student questionnaire items about technical problems encountered. Student login time on the CLIPP cases.</td>
</tr>
<tr>
<td>Major events that compete for M3’s time and attention during their clerkship that could hamper the time and attention they give to CLIPP cases and thus diminish the program’s effectiveness</td>
<td>Student questionnaire items about other major events occurring during this timeframe. Analysis of calendar for major curricular and co-curricular during students’ clerkships</td>
</tr>
</tbody>
</table>
Next, let’s look at the critically important issue of assessing program outcomes. Since the program is being judged primarily on its ability to produce the desired outcomes, it is essential that reliable and valid measures of them are used in the evaluation.

In the CLIPP example, investigators were interested in the short-term program outcomes. Although the EDF grant may only be 1.5 years, it does not necessarily preclude studying medium- and long-term program effects\(^6\). For the CLIPP study, measures of students’ knowledge gains are available through the item database provided with the CLIPP cases. The student satisfaction measures would need to be developed.

**Short-term Outcomes: Dependent Variables**

<table>
<thead>
<tr>
<th>Questions/Constructs</th>
<th>Relevant Information/Measures:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students will have significantly better performance gains on post-tests when they study with CLIPP cases</td>
<td>Pre- and post-test scores on clinical vignettes</td>
</tr>
<tr>
<td>Students will enjoy the self-directed learning process of CLIPP</td>
<td>Student satisfaction ratings</td>
</tr>
</tbody>
</table>

**Inputs**

Since the importance of inputs is primarily for supporting proper implementation of the program, a checklist will likely suffice to document that the resources were available. You will want to describe any instances where these inputs were not available as initially planned and any program adjustments that resulted from the change.

For the CLIPP case, the program inputs were available as planned.

**Assumptions**

Relevant information about assumptions can be gathered from previous research or collected as part of your evaluation. We reviewed the assumptions in our analysis of the CLIPP program design earlier, but as a check for completeness, let’s examine whether we captured information needed to address our assumptions.

- Students understand how to engage in self-directed learning with the CLIPP cases
- CLIPP provides knowledge appropriate level materials for third-year UI medical students
- Students will put forth sufficient time and effort to learn from the CLIPP cases
- Students will select appropriate CLIPP cases, i.e., ones that enable them to address gaps in their knowledge and skills

\(^6\) Typically, longer term outcomes are also more difficult to detect and attribute to program effects. Over time, the number of non-program factors affecting outcomes tends to increase and the initial impact of program factors tends to dissipate. For a discussion of this issue see Norman, G. & Eva, K.W. (2008) Quantitative research methods in medical education. From the Understanding Medical Education Series. Association for the Study of Medical Education. Edinburgh, UK.
Additional items could be added to the student questionnaire to capture information about how students selected and used the CLIPP cases, the effort that they put into the exercises, and whether the case content was appropriate for them. Alternatively, one could structure the program to help ensure that students are selecting CLIPP cases in an appropriate manner, for example, by providing each student with an individualized “prescription” of cases to follow based on pre-test performance. Students’ learning gains for those following the counselor’s advice could be compared to instances where the recommendations were not followed to shed some light on the importance of case selection.

**Relationships**

In the logic model, we described how the program activities are intended to produce the desired outcomes. Now that we have listed some possible measures corresponding to those constructs, we can predict how these measures (variables) should be related if the program is working as intended.

Two concepts are essential for discussing relationships among variables: 1) covariation and 2) causality. If variables $A$ and $B$ covary, changes in the value of $A$ are associated with changes in the value of $B$. Covariation is needed to show that a relationship exists between variables. Thus, for any constructs that our CLIPP logic model posits are related, we should expect to see covariation\(^7\) between their measures.

Although covariation between our program intervention ($A$) and the dependent variable ($B$) is a necessary condition for demonstrating a causal relationship, it is not sufficient. More evidence is required. We also need to establish that:

- $A$ precedes $B$ in time; and
- No third factor, $C$, accounts for the covariation between $A$ and $B$.

The elimination of all possible alternative explanations (i.e., third factors) is a particularly challenging condition to meet in practice. This can be addressed through the research design or through statistical methods of controlling for extraneous variables. Whatever the method used, we are attempting to show what the expected outcome ($B$) for the program participants would have been if they had not undergone the treatment ($A$). In classic experimental designs, a control group is created for this purpose. Since participants are randomly assigned to treatment and control groups, the two groups should be equivalent in all respects (i.e., for all possible third factors, $Cs$) except for their exposure to the treatment.

Examining relationships among variables is central to our question about whether the program worked and subsequent analyses of the factors that enhanced or inhibited its

---

\(^7\) The statistical methods used to analyze covariation depend upon the variables’ levels of measurement and the number and type of variables involved (e.g., two continuous variables – correlation, regression; a categorical independent variable and one continuous dependent variable – t-test, ANOVA; two categorical variables – chi-square, loglinear models).
effectiveness. To illustrate this process, we’ll address these questions for the CLIPP example.

**Does the program work?** Here, we are asking whether the CLIPP activities caused the expected change in the outcomes. To establish that a causal relationship exists, we would need to show:

1) On average, students in the treatment had higher post-test scores (controlling for pre-test scores) than the comparison group (covariance between program activity and outcome);
2) The CLIPP treatment occurred prior to the outcome score (obvious in this case); and
3) No third factor accounts for the relationship between the treatment and outcome. This requires the use of an experimental design or statistical methods that control for alternative explanations.

Possible alternative explanations are often evident in the assumptions and external factors specified in the logic model. For example, we noted that students’ appropriate selection and use of cases is essential. It could be that students with more experience or greater motivation for self-directed learning select the types of CLIPP cases from which they can learn the most. Thus, observed relationships between CLIPP cases and learning gains could be reflective of motivational or experiential differences between the students in the treatment and comparison groups. Alternative explanations like this would need to be eliminated as possibilities if we are to demonstrate a causal relation between CLIPP case usage and learning gains. To do so, we could either use a research design to ensure that treatment and control groups have equivalent levels of experience and motivation for self-directed learning, or we could measure individuals’ level of motivation and experience and control for it statistically in the analysis.

Another way of eliminating alternative explanations is to show that the intervening mechanisms in the CLIPP program operate as posited. Thus, we should expect to see the hypothesized covariation between our intervening mechanisms measures (student satisfaction ratings and item response patterns) and outcome measures in the cases where the treatment is effective. If this more elaborate causal pattern holds, it increases the credibility of our proposed explanation while simultaneously eliminating alternative explanations that could not account for the specified causal chain of relationships.
Now you try.

What are the causal relations you need to test to determine whether your program worked?

What are some plausible alternative explanations for successful program outcomes that you will want to control for?

What intervening mechanisms should covary with the outcomes when the treatment is effective?

If the program did not work, why not?

One possible explanation is that the program actually did work, but the measures and research design were not sensitive enough to detect its effect. This issue can be addressed through statistical power analyses. A power analysis could be conducted when you are setting up your research design to determine how large a sample size you will need to detect a given effect size with a specified statistical test. Additionally, power analyses can be conducted post-mortem to determine how large an effect size your study was capable of detecting.

A second possible explanation is that the program worked, but only for a portion of the participants and/or only under certain conditions. In the CLIPP example, perhaps the program only will work for students who have a certain level of previous experience with self-directed learning. Similarly, perhaps the program only works for certain types of CLIPP cases. Perhaps CLIPP is most helpful when the case concepts are most difficult to apply. We could restrict our analyses of program effectiveness to the appropriate subpopulations or conditions to test these hypotheses. The important point here is to

---

8 There is free software (G*Power) available on the web for conducting such analyses, but it is often helpful to speak with a consultant or colleague knowledgeable on these topics to clarify the statistical and research design issues.
identify these issues **before** you conduct the analysis. The danger of hunting for significant relationships in the data after the fact—known pejoratively as data dredging—is that you will find relationships in the dataset that do not really exist in the population being studied\(^9\).

Still another possibility is that the program didn’t work because it wasn’t properly implemented. We discussed this earlier and will return to it in the final section, when we assemble the components into a final study design.

---

**Now you try.**

Do you have information about the reliability and validity of your outcome measures?

How many subjects do you expect to have in your study?

Is the power of your study a concern?

Might you want to consult with colleagues or a consultant about these issues?

For which subgroups and under what conditions might the effectiveness of the treatment vary?

---

\(^9\) Such relationships are due to random variation, but are not properly screened out as such because of the way the analysis was conducted. The likelihood that you will falsely conclude that random covariation between two variables reflects a statistically significant relationship is actually much greater than the significance level (typically .05) used to conduct the statistical test. The repeated use of significance tests on a dataset raises the likelihood that you will find at least one or more statistically significant relationship. For a fuller discussion of these issues see Millen, B.A. (2005). Multiple analyses in clinical trials: fundamentals for investigators. Statistical Methods in Medical Research, Vol. 14, No. 5, 527-528.
Study Design – Putting these Elements Together

At this point, we have identified possible measures and relationships from which you can build your evaluation. Now we need to consider the best way to put these pieces together to answer your evaluation questions. The Joint Committee on Standards for Educational Evaluation provides some helpful guidelines for designing an evaluation. The standards suggest that the following four central issues be weighed when tailoring your evaluation to the problem at hand.

1. Utility: to ensure that an evaluation will serve the information needs of the intended users.

2. Feasibility: to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

3. Propriety: to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

4. Accuracy: to ensure that an evaluation will reveal and convey technically adequate information about the features that determine worth or merit of the program being evaluated.

In short, the Standards recommend that your evaluation be aimed at producing technically adequate information to serve your evaluation purposes with methods that are ethical and realistic.

Let’s consider these issues as we work through the CLIPP evaluation. It will help us to further refine our evaluation questions and our study design.

First, the purpose of this evaluation is to provide information about the program’s effectiveness to two audiences: the EDF Committee (provided funding), and the program stakeholders. Both audiences are primarily concerned about whether the program worked for UICCOM students and the factors that influenced its effectiveness. The first step in addressing this issue is to determine whether the desired program outcomes were achieved. To do so, it’s essential that there are reliable and valid measures for the learning outcomes. For the CLIPP cases, there is a database with high quality test items. Yet, after reviewing the items, the number of CLIPP cases, and the testing logistics, Tansey and Woodhead were concerned that they may not be able to include a sufficient number of test questions per CLIPP case on the pre- and post-tests to achieve adequate precision for the evaluation study.

---

10 See Appendix B for further discussion of the CLIPP study audiences and their likely evaluation questions.
Another related worry is whether there will be a sufficient number of students to detect a moderate program impact, i.e., an effect size of .5 standard deviations or greater\(^{11}\). From preliminary calculations it appears that a sample size of about 120 would only be sufficient if the students were to be evenly divided between treatment and comparison groups. Yet, practical and ethical reasons militate that students should self-select to treatment and comparison groups rather than being assigned by program faculty/staff. The overall sample size cannot be increased because it is limited to the number of students in the program.

Additionally, the measures of the independent variables have not yet been tested at UI. The statistics gathered via the CLIPP software on login minutes and the proportion of tutorial questions answered seem like they would provide the needed measures of student time and effort spent on the CLIPP cases. Yet, since this is the first time the software would be used by students in the clerkship program, it is possible that it might not work as anticipated.

Given the above limitations and the inevitable implementation glitches that first-time programs encounter, it made most sense to focus this evaluation on aiding proper implementation of the program and ensuring that appropriate measures were in place to detect program impacts. As is often the case with a new program, one cannot develop a definitive test of its effectiveness during initial implementation. Often it takes time to get the program up and running properly and pre-testing of proposed measures before one can properly evaluate whether the program works.

Accordingly, Tansey and Woodhead scaled back the evaluation to focus on the relationship between the independent and dependent variables and any program implementation problems. The following revised questions guided their study.

1) **Original Question**: Did the program work? If not, why not?
**Revised Questions**: Was there a detectable difference between CLIPP and non-CLIPP students’ learning? What problems were there with the measures used for outcomes? Were the sample sizes sufficient to detect CLIPP impacts?

2) **Original Question**: Was the program implemented as intended? If not, why not?
**Revised Questions**: Were the CLIPP usage statistics good measures of program implementation? What problems, if any, were there with these as measures of the independent variables?

3) **Original Question**: How could the program be improved; that is, how could it become more effective and/or efficient?
**Revised Question**: How could the program implementation be improved?

\(^{11}\) Using G*Power, we find that with 128 observations (64 – treatment, 64 – comparison group) we will have an 80% chance (power=.8) of detecting moderate size program effects at a .05 level of significance with an independent samples t-test (two-tailed).
Rather than designing and administering a survey about CLIPP usage, Tansey and Woodhead conducted unstructured interviews with students to gather supplementary information about the program’s implementation and the benefits and challenges students were experiencing.

Now you try.

What are the information needs of the intended users for your evaluation study?

What feasibility considerations impact your evaluation study?

What ethical/propriety issues might affect your evaluation study?

How do issues of accuracy impact your study design?

- Are there questions about the reliability or validity of your measures? If so, how might they be addressed?

- Is there likely to be sufficient statistical power to detect posited effects/relationships? If not, are there ways of increasing the power of the tests?

CLIPP Case Evaluation Design & Analysis

First, let’s look at the questions about detecting program impacts on learning. If the students had been randomly assigned to treatment and control groups one might be able to assume that pre-test score differences between the groups would be negligible. However, remember that “assignment” to group was based on whether a student chose to use a particular case (treatment) or did not (control). Since students selected their CLIPP case topics they may have picked topics areas systematically. For example, they could have selected topics that they understood the least. Any such systematic CLIPP case selection process could mean that students were different in their knowledge prior to using the cases. Consequently, the statistical analysis needs to take into account possible pre-test differences between the groups.
Perhaps the most straightforward approach to incorporate pre-test information would be to calculate difference scores for each student, where

\[
\text{difference score} = \text{post-test score} - \text{pre-test score}.
\]

Unfortunately, there are a number of problems with this approach. Instead, Cook and Campbell (1979) recommend that ANCOVA be used in such situations. This involves using the pre-test score as a covariate to adjust post-test scores and then conducting an ANOVA on the adjusted post-test scores. Tansey and Woodhead used this statistical method and found that although the treatment group had higher adjusted post-test means in seven of the nine topic areas, only one of these seven differences was large enough to be statistically significant.

**ANCOVA Results:**

**CLIPP and Non-CLIPP Posttest Mean Scores adjusted for Pretest Scores**

<table>
<thead>
<tr>
<th>Case Number and Topic</th>
<th>Adjusted Posttest Mean Difference CLIPP–NonCLiPP</th>
<th># of Posttest Questions</th>
<th>F-value for treatment effect</th>
<th>p-value</th>
<th>N_{tut}, N_{comp}</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4: Well-Child Check</td>
<td>-.148</td>
<td>3</td>
<td>1.085</td>
<td>.300</td>
<td>55, 40</td>
</tr>
<tr>
<td>#7: Newborn</td>
<td>.189</td>
<td>5</td>
<td>1.356</td>
<td>.247</td>
<td>42, 54</td>
</tr>
<tr>
<td>#10: Fever</td>
<td>.127</td>
<td>4</td>
<td>.640</td>
<td>.426</td>
<td>45, 47</td>
</tr>
<tr>
<td>#15: Vomiting</td>
<td>.079</td>
<td>6</td>
<td>.129</td>
<td>.721</td>
<td>46, 50</td>
</tr>
<tr>
<td>#16: Abdominal Pain/Vomiting</td>
<td>.335</td>
<td>4</td>
<td>2.466</td>
<td>.120</td>
<td>24, 70</td>
</tr>
<tr>
<td>#21: Rash</td>
<td>.357</td>
<td>2</td>
<td>12.805</td>
<td>.001</td>
<td>60, 36</td>
</tr>
<tr>
<td>#23: Lethargy and Fever</td>
<td>.054</td>
<td>1</td>
<td>.393</td>
<td>.532</td>
<td>32, 64</td>
</tr>
<tr>
<td>#26: Not Gaining Weight</td>
<td>-.041</td>
<td>2</td>
<td>.099</td>
<td>.753</td>
<td>34, 63</td>
</tr>
<tr>
<td>#30: Sickle Cell Disease</td>
<td>.141</td>
<td>3</td>
<td>.868</td>
<td>.354</td>
<td>41, 55</td>
</tr>
</tbody>
</table>

The next step would be to do a power analysis to determine the effect sizes that this study was capable of detecting. If this study was not sufficiently powered to detect appropriate effect sizes, adjustments could be made to the measures and effective sample sizes to increase the statistical power in subsequent studies.

Now, let’s look at the second set of revised evaluation questions. Were the CLIPP usage statistics good measures of program implementation? What problems, if any, were there with the measures used for the independent variables? The results from the study, based on data from 120 students were:

- Students averaged 734 minutes on CLIPP cases throughout the clerkship  
  - (time represents duration of time that the web site was connected, not necessarily direct learning)
- Averaged over the year, each student completed about 12 cases  
  - (>75% of case questions answered)
These measures of time and effort were not as clean as anticipated. “Login time” recorded all the time students were logged onto the application, even when there were long periods of inactivity. Thus, if students forgot to sign out after working on the case, the inactive login minutes of computer connection were also included in the time count. Tansey and Woodhead also learned that “the proportion of case questions” answered might be a problematic measure of students’ “effort.” Some students reported guessing at the case questions without doing the reading so that they could gain access to the case explanations. Consequently, refinements need to be made to these processes before the measures can be used in the analysis.

Lastly, let’s look at how the program implementation could be improved. Answering this question requires pulling the findings together into some general conclusions about the initial implementation of the program. To do so, Tansey and Woodhead reviewed the data discussed above, analyzed the CLIPP curriculum, and asked students about their impressions of the program. They found that there is great potential for the CLIPP program and noted that

- CLIPP was a well utilized resource throughout the Pediatric Clerkship
- The CLIPP cases have been designed to reflect the curriculum of the clerkship.

Yet, accurate assessment of its affect on student learning will require some retooling of the measures for the outcomes and independent variables. With the information provided by their preliminary study, they are now in an excellent position to advance their program evaluation effort and improve the program’s effectiveness.

**Summary**

The evaluation method we have illustrated in this workbook centers around the use of a logic model to develop a thorough understanding of how the program should work in theory. Based on this theory, hypothesized relationships can be developed among empirical measures to show how the program is expected to work in practice. Which relationships you choose to study and the methods you use, depend upon a number of factors (e.g., maturity of the program, availability of reliable and valid measures for the program variables, uses of the evaluation information). To aid with decisions about the appropriate size, scope, and focus of your evaluation, it was recommended that you use the program evaluation standards of utility, feasibility, propriety, and accuracy.
APPENDIX A
CLIPP Case Summary

Computer-Assisted Learning in Pediatrics Project (CLIPP)

The directors of a pediatrics clerkship wished to evaluate the usefulness of a case bank of simulated cases that were developed by their national clerkship organization. They wondered how best to use these cases within their clerkship. The following is a brief description of their project. Using this information, you can go through the steps we have outlined in this manual. “Expert” approaches to designing an evaluation can be found later in this manual and used to compare to yours.

Purpose of the project:
To determine whether self-directed computer-based cases are an effective tool to help 3rd-year medical students correct knowledge deficiencies. The project will use cases developed through a multi-institutional project that have undergone peer and student review, however, models of how to incorporate the cases within other clerkship activities have not been evaluated. In this model, the cases will be used as an adjunct to other learning activities.

Target Audience:
3rd-year medical students rotating through a required Pediatrics Clerkship.

Activities:
Self-instructional cases:
This web-based program has 31 interactive virtual patient cases designed to cover all the core content of a pediatrics clerkship as defined by the Council on Medical Student Education in Pediatrics (COMSEP). The cases are intended to supplement traditional clerkship didactics and patient care activities.

The cases were developed through a 3-year multi-institutional development process involving clerkship directors from 30 programs.* They are in use by more than 80 medical schools across the U.S. and Canada. They received an innovation award for the Northeast Group on Educational Affairs (an Association of American Medical Colleges regional group).

The cases are interactive and make use of multimedia. Students are required to make decisions about diagnoses and clinical management at key points in the cases and to justify their decisions based on the findings in the case. They receive feedback.

Instructional Model (how the cases will be used):
All students (approximately 140) rotate through the clerkship in groups of 16-18 per 6-week block. On the first day of the clerkship we will administer a 30-question multiple choice test consisting of questions that include brief clinical vignettes related to 15 of the case topics. The questions will be taken from the COMSEP national test item bank that has item analysis statistics and is matched to the core curriculum and thus the cases. Each case will have 4 test items, with 2 randomly selected for the pretest and 2 for the post test. On the last day of the clerkship we will administer a post-test that includes questions on the 15 content areas. The pre-test will be required but not counted towards their grade.

Based on pre-test performance each student will receive a personalized “prescription” of recommended CLIPP cases. Students will be allowed open access to all 31 cases at any time during the 6-week

* See supplement for additional information about how the cases were developed.
clerkship. Students are asked to complete at least 6 cases. Each case takes about 45 minutes to complete.

**Time Schedule:**
The program will run for one complete academic year. Data will be collected every clerkship period (6 weeks) for the year and analyzed quarterly.

**Outcomes:**
This project will allow us to determine whether the cases are an effective tool to correct knowledge deficiencies in 3rd-year medical students on this rotation. It will also examine how students are self-directed in their learning with the cases. Finally it will provide data on the knowledge level of incoming students.

**Evaluation Plan**
We will compare pre and post test scores on the 15 content areas. We will correlate extent of use of the cases (the number of cases completed and the specific cases accessed for each student) with knowledge gain. We will also look at timing of case usage across the clerkship and time spent on cases compared to national data about mean time.

In the analysis, students will be divided into two main groups based on whether they engaged in a particular case (EC) or did not (NC). Across subsets of items, pre-test and post-test results will be used to calculate gain scores as follows:

(ECpost-ECpre) = Gain
(NCpost-NCpre) = gain

Each student will have a subset of results included in both the EC and NC groups. Each student serves as his or her own control and is pseudo-randomly assigned to a group based on the cases used. We will conduct a check on the success of the pseudo-random group assignment by checking the equivalence of pretest scores in the two groups. If they are not equivalent we will use ANCOVA.

The final analysis will be to conduct a t-test for the gain scores.

---

Supplemental information: Development and initial evaluation of the computer cases

The cases were developed by the Council on Medical Student Education in Pediatrics (COMSEP) through federal grant over a three-year period. The goal was to develop quality computer-assisted instructional material to teach the COMSEP core curricular objectives. This was accomplished through a multi-institutional team approach that paired pediatrics clerkship directors with clerkship directors experienced in computer-assisted curriculum development. A content team ensured that the series contained all the core COMSEP objectives but were not redundant, and that scenarios were realistic. A development team created a standard case template. Cases were evaluated using blinded peer review and student pilot-testing at three COMSEP schools. Results of the peer and student reviews were published in 2005 (Academic Medicine 80:9:847-855, 2005) and provided validation of the appropriateness and accuracy of the case content to the COMSEP objectives and the usefulness of the case structure, based on peer and student perspectives. The study did not look at learning effectiveness outcomes.

This case is based on a project by Michal Tansey and Jerold Woodhead, Pediatrics. Some information was modified for the purpose of this case illustration.
APPENDIX B
Sample Method for Developing and Prioritizing Evaluation Questions

To clarify your study purpose and research questions, it is often helpful to begin by thinking about your intended audiences and their information needs. In the context of the EDF program, you will probably be interested in three main audiences: 1) EDF Committee - the grant sponsor; 2) Program Implementers/Stakeholders – you and your colleagues; and 3) Professional Audiences for the publication of your findings. These audiences will have somewhat different interests. The EDF Committee will want a report describing the activities that were completed and whether these actions produced the desired outcomes. You and your colleagues will likely want to regularly track the program activities that have been implemented, have some early indicators of how it appears to be working, and periodically assess what you can do to make the program more effective in producing the desired outcomes. Professional audiences will likely be concerned about the credibility of the information and its generalizability to other settings.

As described in the above situation, there are typically a few different purposes to be addressed in such studies that naturally lead to some general types of evaluation questions: 1) accountability – was the project implemented as described in the grant? 2) summative evaluation – did the project work? What are the program’s benefits relative to its costs? Overall, how worthwhile or valuable is the program? and 3) formative evaluation – how could the project be improved and made more efficient or effective?

To translate these into more specific questions, we would use the program analysis contained in the logic model to create specific research/evaluation questions. For example, let’s consider the car buying example. A fundamental assumption in the program is that engaging in an objective, systematic decision-making approach will result in a more satisfactory purchase. An accountability question would be: “Did we engage in a systematic process, as described in the logic model, or did emotion or car salesman persuasiveness lead us in intended directions? A summative question is whether we are satisfied with the car, initially and five or more years later. We would need to qualify what we mean by satisfying in terms of cost, reliability, drivability, etc. The summative questions might also include decisions about whether to keep the car or get rid of it. Formative questions could include how to improve the decision-making process in the future.

Now consider the CLIPP case example. Assuming we have the above audiences and purposes for the study (accountability, summative, formative) what specific questions should we address in the evaluation? As a starting point, we could use the following chart to help us think about the audiences for this project and the types of information that they might need. This will help us look at the program from their viewpoint. You are unlikely to address all potential questions but the analysis will help you prioritize based on importance, commonality of interest across audiences, and potential impact of the answer.
## CLIPP Case Illustration: Sample Information Needs for Audiences

<table>
<thead>
<tr>
<th>Audiences:</th>
<th>Accountability questions</th>
<th>Summative questions</th>
<th>Formative Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Funding source:</strong> (e.g., EDF Committee)</td>
<td>Was the program implemented as intended? Why or why not and what was the impact?</td>
<td>Is the program effective?</td>
<td>Will the program be maintained? Why or why not?</td>
</tr>
<tr>
<td><strong>2. Stakeholders:</strong></td>
<td><strong>Clerkship director</strong> Was the program implemented as intended, e.g., Did the pre-test perform as intended? Did all students take it? Did students use the cases as intended? Why or why not?</td>
<td>Is the program effective? Learning effectiveness and efficiency?</td>
<td>Is the program being implemented as intended?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the cost to conduct the program?</td>
<td>What would improve the program?</td>
</tr>
<tr>
<td>Teaching faculty</td>
<td></td>
<td>Is the program effective?</td>
<td>How does it affect my teaching?</td>
</tr>
<tr>
<td>Students</td>
<td></td>
<td>Will time invested in using these cases lead to important learning?</td>
<td></td>
</tr>
<tr>
<td>Department chair</td>
<td></td>
<td>Is the program effective?</td>
<td>What is the cost to administer the program?</td>
</tr>
<tr>
<td>Dean</td>
<td></td>
<td>Same as above related to dissemination to other clerkships as a model approach</td>
<td></td>
</tr>
<tr>
<td>Other clerkship programs locally or nationally</td>
<td>Same as above related to dissemination as a model approach</td>
<td></td>
<td>What components of the program are most important to assure effectiveness?</td>
</tr>
<tr>
<td><strong>3. Professional Audiences</strong> (e.g., medical education Journals; professional meetings)</td>
<td>Are the claims regarding program effectiveness supported by sufficiently credible information?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
With an information needs matrix, like the above, and a logic model, you have the essential elements for developing, refining, and prioritizing your study questions. Using the process discussed earlier in Step 2 of the evaluation process (pp. 7-12), you can use these two documents to flesh out your program description and refine your initial questions listed in the information needs matrix.

Prioritizing research and evaluation questions

Using this method to analyze your program information needs may result in more questions than you can address with the resources you have. You may need to prioritize. These questions can help guide those decisions.

- Will the answer to the question have an impact on decisions?
- Will the answer to the question reduce uncertainty? (Will it tell you what to do to improve the program?)
- Will the evaluation question be of interest to key stakeholders? (yourself, participants, administrators, funding agencies, etc.)
- Is the question answerable? (possible and feasible)
- Does it merit the expense of time and resources required to answer it?
- Will it explain the program to others who might want to replicate it?
The Kirkpatrick Four-Level program evaluation model is well-known in the training and business fields. While sometimes criticized as being simplistic, its simplicity makes it readily understood and adaptable to faculty development programs.

The fundamental assumptions and principles of this model are:

- Program effectiveness can be measured at 4 levels. These include participants’ reactions to the program; their learning from the program; their behavior on the job as a result of learning; and the results that occurred (impact).
- The levels are sequential, with higher levels dependent on achievement at lower levels. Learning is less likely to occur if the program is poorly received. Likewise, if participants don’t learn anything from the program, they won’t change their work behavior.
- Always measure at level one and build the other levels in sequence, without skipping a level. This rule is based on the fact that it is difficult to prove beyond doubt that a program produces results because there are too many variables that you cannot control. You can, however, look for a “chain of evidence” leading from the program through the four levels. This concept of causal sequence is similar to another evaluation concept called “program logic,” which is a description or hypothesis of how a program works to produce particular results (see glossary).

This model is illustrated in the diagram below showing the chain of events.

The intervening factors are not under the trainers’ control but can support or block changes in behavior in practice. These factors increase in number the further the “link” is from the program. In particular, behavior change can be influenced by factors that

- Enable one to perform the behavior on the job (or prevent one from performing). These include things within the organization that facilitate or impede new behavior including social interactions and technical systems (work processes, policies, resources).
- Reinforce the behavior once it occurs (or extinguish it). These include characteristics of the organization that reward or punish certain behaviors

Kirkpatrick recommends that evaluators set acceptable standards at each level they will measure, prior to collecting data. Envision how you hope the program will be implemented and acceptable reactions from participants, what they really must learn from the training event, what teaching behavior you hope faculty will engage in, and how those behaviors will affect learners and the teaching enterprise.
Evaluators of faculty development programs must decide from the onset, at what level they will evaluate. Kirkpatrick recommends that all programs be evaluated at Level 1 because it is easy to do and is the prerequisite to accomplishing the other levels. Evaluators may have a need to show further evidence of a program’s effectiveness, in which case they should build each level sequentially. Kirkpatrick also proposes a “0” level which is an assessment of current benchmarks for performance (before program implementation). In other words data collected during the initial needs assessment, can be used for comparison after implementing or revising the program.

Further description of the levels follows, along with suggestions for how to measure each level.

**Level 1: Reactions**

What it is: Participants’ perceptions and opinions about the program generally but also specific questions about subcomponents: content, instructor, audiovisuals, handouts, facilities, etc. Consider participant’s reactions to components of your program that you have deliberately included so as to ensure they can transfer the new learning to practice, such as use of realistic examples, practice opportunities, and assistance at understanding the importance of, and application to, their work. Level 1 assessment gives you an idea of how motivated participants are and provides important formative evaluation data that you can use to revise your program if you offer it repeatedly.

How to do: Usually collected immediately after the program using questionnaires with items that ask participants to rate the quality of, or their satisfaction with, the program. Also includes open-ended items to help you identify strengths, weaknesses, suggestions for improvement.

**Level 2: Learning**

What it is: Changes in participants’ knowledge, skills, and attitudes are the focus here; therefore this level is concerned with measuring whether the learning objectives have been achieved as a result of your training program.

How to do: Kirkpatrick recommends assessing learning immediately after the program so you will know if some objectives have not been met. Knowledge can be measured with quizzes or exams, skills preferably with performance tests, and attitudes can be measured with questionnaires.

**Level 3: Behavior Transfer**

What it is: Changes to behavior in practice—“on the job.” Do they use the knowledge and skill learned to change the way they practice? Kirkpatrick argues that in order to interpret evaluation results at this level, you must also measure the prior two levels, for if there are no behavior changes you will want to know whether they learned anything to change.

How to do: Behavior in practice can be measured in many ways including self-assessment but preferably through multiple sources of information. Assessment may be made by sampling specific instances, (e.g., observation by a trained rater), or generalized across many encounters. A pre-assessment is desirable but if you can’t do that an alternative is the retrospective pre-post assessment, in which you ask trainees, after training, to rate how they were before training.

- Self-rating of practice behavior (e.g., Likert-type scales)
- Products of their work
- Surveys and questionnaires
- Observations by peers, supervisors, patients/clients, or trained evaluators using checklists or rating scales
- Interviews and focus groups

**Level 4: Results (Impact)**

What it is: Longer-range effects of the program on participants, on those who interact with them (e.g., patients and their families) or on the organization as a whole.

How to do: Kirkpatrick recommends that you identify performance standards early in program development—predict or hypothesize what effect the desired changes in behavior could have on others and the organization, as well as some potential undesirable effects. This will help minimize bias in what you choose to measure.

Impact is difficult to evaluate because training variables cannot be separated from intervening variables. It will be difficult to measure the impact of a single specific skill (e.g., giving feedback). You may have greater success in measuring the impact of a program that addresses multiple related skills, a broader competency. “Proof” will be difficult to establish, but you may find a “preponderance of evidence.” Evaluation at this level benefits from a controlled research design, so you must consider the costs versus the need for this evidence.

The timing for assessing levels 3 and 4 is important and may occur weeks or months after the program, depending on the nature of what was learned. You need to allow enough time for new behaviors to be incorporated and have an impact. You may want to repeat the evaluation.

In addition to measuring behavior and results at these two levels, you can also attempt to measure characteristics of the environment (organizational culture and climate) that can affect behavior and resulting impact. If you do this before implementing the program you might be able to develop strategies to accommodate the environment to which your trainees will return. Factors that enable and reinforce changes in practice include:

- Have they had opportunity to use the new knowledge or skill?
- What factors in their organization support use of the behaviors they have learned?
- What sorts of encouragement or rewards do they receive? (extrinsic motivation)
- How well do they like the new behavior? Do they think it’s useful, satisfactory, or rewarding? (intrinsic motivation)

Finally, you may also need to document other events or activities that occurred during this timeframe that may have impacted the training or subsequent behavior.

APPENDIX D
Reviewing Quantitative Educational Research/Evaluation Studies

Here are some key pieces of information that you will want to consider. For #6-8 below, it is often helpful to consult with colleagues who have expertise in the particular measures, statistical methods, or research designs used in such research.

1. Research Questions: List the research questions that the study seeks to answer.

2. Study Population/Sample: Who are the subjects in the study? Capture any descriptive information about the subjects.
   - Do the research questions pertain only to the subjects in the study?
   - Are the study subjects intended to represent a larger population? If so, what sampling methods were used to draw the subjects from the population?
   - What percentage of the subjects participated in the study? Did any subjects withdraw from the study before it was completed?

3. Applicability to your Research Questions and Study. Are the research questions and study population/sample similar enough to your research situation to merit further study of the article? If yes, proceed to #4.

4. Key Constructs & Assumptions. Write down the definitions of the key constructs that are related to the research questions. Do the researchers make any assumptions about the study subjects, program elements or context, that are not captured by the key constructs? If so, list them.

5. Relations among the Key Constructs. Draw a diagram showing the expected relationships among these constructs and assumptions. If the study is a program evaluation, draw a logic model using the attached form to specify the types of variables, assumptions, and the interrelations among them.

6. Measures of Key Constructs. Identify the measures (operational definitions) used for the key constructs. Record any evidence presented regarding the reliability and validity of the measures. List the measures and note any concerns you have about how well they reflect the constructs discussed in #4.

7. Relations among the Measured Variables. Do the relationships described in #5 appear as expected among the measured variables? What statistical tests were performed to examine the relationships? How many observations were used in the tests? Ability to detect relations among the empirical measures can be affected by a variety of threats to statistical conclusion validity like the following:
   - Low statistical power for the tests
   - Violated assumptions of statistical tests
   - High levels of error – variation in treatment environment, variation in the way the treatment was administered, unreliable measures, and heterogeneous subjects

Do you see any of these possible threats in the study? If so, record them.

8. Evaluate the Evidence and its Applicability to your Research. Were any of the relations detected in the study interpreted as causal? If so, did their research design and analysis support causal inference? To which student groups/populations do their findings apply? How well did their study address the research questions? Overall, how credible was the evidence they used to answer the questions? What limitations does the study have?

---