2

An Introduction to Evidence-Centered Design

Although assessment design is an important part of this book, we do not tackle it in a formal way until Part III. Part I builds up a class of mathematical models for scoring an assessment, and Part II discusses how the mathematical models can be refined with data. Although throughout the book there are references to cognitive processes that the probability distributions model, the full discussion of assessment design follows the discussion of the more mathematical issues.

This presents two problems. First, a meaningful discussion of the statistical modeling of the assessment requires a basic understanding of the constraints and affordances of the assessment design process. The second is that the discussion of the statistical models and processes requires certain technical terms, in particular, *proficiency model, evidence model, task model,* and *assembly model,* that are not formally defined until Chap. 12. This chapter provides brief working definitions which will be sufficient to describe the mathematical models, leaving the more nuanced discussion of assessment design until after the mathematical tools have been defined.

Evidence-centered design (*ECD*) is an approach to constructing educational assessments in terms of evidentiary arguments. This chapter introduces the basic ideas of ECD, including some of the terminology and models that have been developed to implement the approach. In particular, it presents the high-level models of the Conceptual Assessment Framework (see also Chap. 12) and the four-process architecture for assessment delivery systems (see also Chap. 13). Special attention is given to the roles of probability-based reasoning in accumulating evidence across task performances, in terms of belief about unobservable variables that characterize the knowledge, skills, and/or abilities of students. This is the role traditionally associated with psychometric models, such as item response theory and latent class models. Later chapters will develop Bayesian network models which unify the ideas and provide a foundation for extending probability-based reasoning in assessment applications more broadly. This brief overview of evidence-centered design,

R. G. Almond et al., *Bayesian Networks in Educational Assessment*,
Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4939-2125-6_2

then, provides context for where and how graphical models fit into the larger enterprise of educational and psychological assessment.

2.1 Overview

All educational assessments have in common the desire to reason from particular things students say, do, or make, to inferences about what they know or can do more broadly. Over the past century a number of assessment methods have evolved for addressing this problem in a principled and systematic manner. The measurement models of classical test theory and, more recently, item response theory (IRT), latent class analysis, and cognitive diagnosis modeling, have proved quite satisfactory for the large scale tests and classroom quizzes with which every reader is by now quite familiar.

But off-the-shelf assessments and standardized tests are increasingly unsatisfactory for guiding learning and evaluating students' progress. Advances in cognitive and instructional sciences stretch our expectations about the kinds of knowledge and skills we want to develop in students, and the kinds of observations we need to evidence them (Pelligrino et al. 2001; Moss et al. 2008). Advances in technology make it possible to evoke evidence of knowledge more broadly conceived, and to capture more complex performances. One of the most serious bottlenecks we face, however, is making sense of complex data that result.

Fortunately, advances in evidentiary reasoning (Schum 1994) and in statistical modeling (Gelman et al. 2013a) allow us to bring probability-based reasoning to bear on the problems of modeling and uncertainty that arise naturally in all assessments. These advances extend the principles upon which familiar test theory is grounded to more varied and complex inferences from more complex data (Mislevy 1994).

We cannot simply construct "good tasks" in isolation, however, and hope that someone else down the line will figure out "how to score them." We must design a complex assessment from the very start around the inferences we want to make, the observations we need to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them (Messick 1994). We can expect iteration and refinement as we learn, from data, whether the patterns we observe accord with our theories and our expectations; we may circle back to improve our theories, our tasks, or our analytic models (Mislevy et al. 2012). But the point is that while more complex statistical models may indeed be required, they should evolve from the substance of the assessment problem, jointly with the purposes of the assessment and the design of tasks to provide observable evidence.

ECD lays out a conceptual design framework for the elements of a coherent assessment, at a level of generality that supports a broad range of assessment types, from familiar standardized tests and classroom quizzes, to coached practice systems and simulation-based assessments, to portfolios and student–tutor interaction. The design framework is based on the principles of evidentiary reasoning and the exigencies of assessment production and delivery. Designing assessment products in such a framework ensures that the way in which evidence is gathered and interpreted bears on the underlying knowledge and the purposes the assessment is intended to address. The common design architecture further aids coordination among the work of different specialists, such as subject matter experts, statisticians, instructors, task authors, delivery-process developers, and interface designers. While the primary focus of the current volume is building, fitting, testing, and reasoning with statistical models, this short chapter places such models into the context of the assessment enterprise. It will serve to motivate, we hope, the following chapters on technical issues of this sort. After that machinery has been developed, Chap. 12 returns to ECD, to examine it more closely and work through some examples.

Section 2.4 describes a set of models called the *Conceptual Assessment Framework*, or *CAF*, and the *four-process architecture* for assessment delivery systems. The CAF is not itself the assessment design process, but rather the end product of the assessment design process. Although this book does not cover the earlier stages of the design process, Sect. 2.3 touches on them briefly. Mislevy, Steinberg, and Almond (2003b) present a fuller treatment of ECD including connections to the philosophy of argument and discussions of the earlier stages of design. Almond et al. (2002a) and Almond et al. (2002b) amplify the delivery system architecture and its connection to the design.

One of the great strengths of evidence-centered design is that it provides a set of first principles, based on evidentiary reasoning, for answering questions about assessment design. Section 2.2 provides a rationale for assessment as a special case of evidentiary reasoning, with validity as the grounds for the inferences drawn from assessment data (Cronbach 1989; Embretson 1983; Kane 1992; Kane 2006; Messick 1989; Messick 1994; Mislevy 2009). ECD provides a structural framework for parsing and developing assessments from this perspective.

2.2 Assessment as Evidentiary Argument

Advances in cognitive psychology deepen our understanding of how students gain and use knowledge. Advances in technology make it possible to capture more complex performances in assessment settings, by including, for example, simulation, interactivity, collaboration, and constructed responses in digital form. Automated methods have become available for parsing complex work products and identifying educationally meaningful features of them Williamson et al. (2006b).

The challenge is in knowing just how to put all this new knowledge to work to best serve the purposes of an assessment. Familiar practices for designing and analyzing single-score tests composed of familiar items are useful because they are coherent, but the schemas are limited to the constraints under which they evolved—the kinds of tasks, purposes, psychological assumptions, cost expectations, and so on that define the space of tests they produce. Breaking beyond the constraints requires not only the means for doing so (through advances such as those mentioned above) but schemas for producing assessments that are again coherent, but in a larger design space; that is, assessments that may indeed gather complex data to ground inferences about complex proficiency models, to gauge multidimensional learning or to evaluate multifaceted programs—but which are built on a sound chain of reasoning from what we propose to observe to what we want to infer. We want to design in reverse direction: What do we want to infer? What then must we observe in what kinds of situations, and how are the observations interpreted as evidence?

Recent work on validity in assessment lays the conceptual groundwork for such an approach. The contemporary view focuses on the support—conceptual, substantive, and statistical—that assessment data provide for inferences or actions (Messick 1989). From this view, an assessment is a special case of evidentiary reasoning. Messick (1994) lays out the general form of an assessment design argument in the quotation below. (We will look more closely at assessment arguments in Sect. 12.1.2.)

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics (p. 17).

This perspective organizes thinking for designing assessments for all kinds of purposes, using all kinds of data, task types, scoring methods, and statistical models. An assessment interpretation reasons from what we observe to what we then believe about students' proficiencies. Assessment design reasons in the reverse direction, laying out the elements of an assessment in a way that will support the needed interpretations.

For the purpose of the assessment, what are the proficiencies we are interested in? In what situations do people draw on them, to accomplish what ends, using what tools and representations, and producing what kinds of outcomes? Taking context and resources into account, we consider task situations we can devise and observations we can make to best ground our inferences. If interactions are key to getting evidence about some proficiency, for example, we can delve into what features a simulation must contain, and what the student must be be able to do, in order to exhibit the knowledge and skills we care about. We craft scoring methods to pick up the clues that will then be present in performances. We construct statistical models that will synthesize evidence across multiple aspects of a given task performance, and across multiple task performances. These decisions in the assessment design process build the inferential pathway we then follow back from examinees' behaviors in the task setting to inferences about what they know or can do. From an evidentiary reasoning perspective, we can examine the impact of these design decisions on the inferences we ultimately want to make.

As powerful as it is in organizing thinking, simply having this conceptual point of view is not as helpful as it could be in carrying out the actual work of designing and implementing assessments. A more structured framework is needed to provide common terminology and design objects that make the design of an assessment explicit and link the elements of the design to the processes that must be carried out in an operational assessment. Such a framework not only makes the underlying evidentiary structure of an assessment more explicit, but it makes it easier to reuse and to share the operational elements of an assessment. The evidence-centered design models address this need.

2.3 The Process of Design

The first step in an assessment design is to establish the *purpose* of the assessment. Many fundamental design trade-offs, e.g., assessment length versus reliability, breadth across multiple aspects of proficiency versus depth in a single proficiency, are ultimately resolved by deciding how to best meet the purpose of the assessment. Fixing the purpose of the assessment early in the process has a marvelous focusing effect on the design and development processes.

Fixing the purpose, however, is easier said than done. Different test users may have different and competing purposes in mind for a proposed assessment. Expectations can be unrealistic, and can change over time. The purpose of an assessment often starts as somewhat vague in the beginning of the design process and becomes further refined as time goes on.

The ECD framework describes the assessment design process in three stages: *domain analysis*—gathering and organizing information related to the cognitive background of the assessment as well as the purposes and constraints of the design process; *domain modeling*—building a preliminary sketch of the assessment argument as a general, reusable framework for a family of possible assessments; and the *conceptual assessment framework*—filling in the details of the initial sketch, particularly resolving design decisions to focus the product on a particular purpose.

The lines between requirements-gathering, analysis, design, and implementation are difficult to draw (indeed, the authors have argued among themselves about which of the steps of the ECD process correspond to which steps of the general engineering workflow). Describing the ECD process in phases might seem to suggest a waterfall development process, where each stage flows into the next and the flow is just one way. Real-world assessment design processes are usually iterative, with prototypes and cycles; things learned at later stages of design often prompt the designer to revisit, rethink, and revise work done at the earlier stages. Mislevy, Steinberg, and Almond (2003b) discussed the ECD design process in more detail.

For the most part, this book does not delve deeply into these design issues so that it can focus on the theory, the roles, and the mechanics of Bayesian networks in the assessment argument. Most of the examples assume that the conceptual assessment framework has already been specified. Only with the Biomass example of Chaps. 14 and 15 do we work through the design process from the very beginning: from targeted educational standards, through the CAF, to the innovative, interactive tasks, and a Bayes nets scoring model that result from the unified design process. It does not hurt to say again, though, that complex measurement models such as Bayesian networks will provide the greatest value when they arise from a principled design process to serve an evidentiary argument, rather than applied retrospectively to data that are collected without clear hypotheses connecting proficiencies and the situations and performances that reveal them (iteration and refinement notwithstanding).

Section 2.4, then, describes the basic design objects of the CAF. The domain-model design objects are basically lighter weight versions of their CAF counterparts; detailed enough to support the assessment argument, but not yet detailed enough to support implementation. In the domain modeling phase, the design team are encouraged to think about how the assessment argument would play out for multiple purposes and in multiple settings. It helps to identify opportunities in which argument structures from one assessment can be reused in another.

One kind of design object, developed in the early stages of the design process but used extensively in the CAF, is the *claim*. A claim is a statement about a participant that the assessment will provide evidence for (or against). Claims are important because they give clarity to the purpose of an assessment. One of the most important design decisions is deciding which claims will be the primary focus of an assessment. Indeed, the whole question of validity could be framed as determining to what extent an assessment really supports its claims.

A simple example, used through the rest of the chapter, illustrates these ideas.

Example 2.1 (Calculus Placement Exam). *University C requires all students to take 2 years of calculus, in the form of a two-semester freshman sequence followed by a two-semester sophomore sequence. Typically a student starts with the first semester in the freshman year, but some students (particularly those who took an advanced calculus class in high school) start with the second semester, or with the third semester with the sophomore cal-*

culus class. Some students do not have the necessary background to begin the sequence, and should take a precalculus remedial course first. University C administers a placement exam to all incoming freshmen to determine how to best place them into the calculus sequence.

Claims in this assessment are based on the student having proficiencies that are addressed in each of the courses in the calculus series. Examples include, "Student can integrate functions of one variable," and "Student can find partial derivatives of multivariate functions." Note that there may be competing interest in the claims. For example, the Physics department may have more interest in the claim "Student can solve integrals in two and three dimensions" while the Math department is more interested in the claim "Student can construct a valid mathematical proof."

Often claims are arranged hierarchically. For example, the claim "Student can integrate functions of one variable" involves the subclaims, "Student can integrate polynomial functions" and "Students can integrate trigonometric functions" as well as the subclaims "Student can use transformation of variables to solve integrals" and "Student can use partial fractions to solve integrals." "Student can construct a valid mathematical proof" will need further specification with respect to the particular models and the kind of the proof at issue (e.g., existence proof, induction, construction, proof by contradiction). It will be seen that a set of claims is not sufficient to determine the proficiency model for a given purpose. Composite claims that bundle finer-grained claims dealing with skills in the same semester are good enough for course placement, but the finer-grained claims would be distinguished for quizzes and diagnostic tests during a semester.

In this particular case, the claims are relatively easy to establish. They will fall naturally out of the syllabus for the calculus series and the calculus text books. They are not simply a list of topics, but rather the kinds of problems, proofs, and applications a student is expected to be able to carry out.

Another frequent source of claims is the educational standards published by states and content area associations, such as the *Next Generation Science Standards* (NGSS Lead States 2013). Grain size and specificity vary from one set of standards to another, and often they need to be refined or clarified to take the form of claims. They may not be phrased in terms of targeted capabilities of students, or indicate what kinds of evidence is needed. It is not enough say, for example, that "Student understands what constitutes a valid mathematical proof." Chapter 14 provides an example of moving from standards to a framework of claims to ground an assessment.

Claims play two key roles in domain modeling: (1) including and excluding specific claims clarifies the purpose of the assessment, and (2) laying them out starts the process of developing an assessment argument. These roles are so important that while most domain modeling design objects are refined and expanded in the CAF, claims remain largely in their initial form.

2.4 Basic ECD Structures

In an ECD, an assessment design is expressed through a collection of objects called the CAF. In any particular assessment, the objects in the CAF models described in general terms in Sect. 2.4.1 will need to have been designed to address the purposes of that particular assessment. In line with the Messick quotation cited above, the characteristics of tasks have been selected to provide the opportunity to get evidence about the targeted knowledge and skills (i.e., the claims); the scoring procedures are designed to capture, in terms of observable variables, the features of student work that are relevant as evidence to that end; and the characteristics of students reflected as proficiency variables summarize evidence about the relevant knowledge and skills from a perspective and at a grain size that suit the purpose of the assessment. The CAF models provide the technical detail required for implementation: specifications, operational requirements, statistical models, details of rubrics, and so on.

CAF models provide specifications, but specifications are not an assessment. As examinees and users of assessment ourselves, we see activities: Tasks being administered, for example, and students interacting with task contexts to produce essays or solve problems, raters evaluating performances or automated algorithms evaluating work, score reports being generated, and feedback being given to students in practice tests. We will organize all of this activity in terms of *processes*, as described below. It is the CAF that specifies the *structure* and the *relationships* of the all content, messages, and products involved in the processes. In other words, the CAF lays out the structural elements of an assessment that embody an assessment argument. The delivery processes described below bring the assessment to life. They are real-world activities that interact with students, gather evidence, and support inference using those structures.

In describing both the design and implementation of scoring models and algorithms, it is useful to have a generic model of the assessment delivery process. Section 2.4.2 describes the four-process architecture that forms a reference model for the delivery of an assessment. The four processes of the delivery system carry out, examinee by examinee, the functions of selecting and administering tasks, interacting as required with the examinee to present materials and capture work products, then evaluating responses from each task and accumulating evidence across them. The information in the CAF models specs out details of the objects, the processes, and the messages that are all interacting when an assessment is actually in play. Any real assessment must have elements that correspond to the four processes in some way. Thus, exploring how assessment ideas play out in the four process framework provides an understanding about how they will play out in specific assessment implementations.

2.4.1 The Conceptual Assessment Framework

The blueprint for an assessment is called the CAF. To make it easier to rearrange the pieces of the framework (and deal with them one at a time when appropriate), the framework is broken up into pieces called *models*. Each model provides specifications that answer such critical questions as "What are we measuring?" or "How do we measure it?"

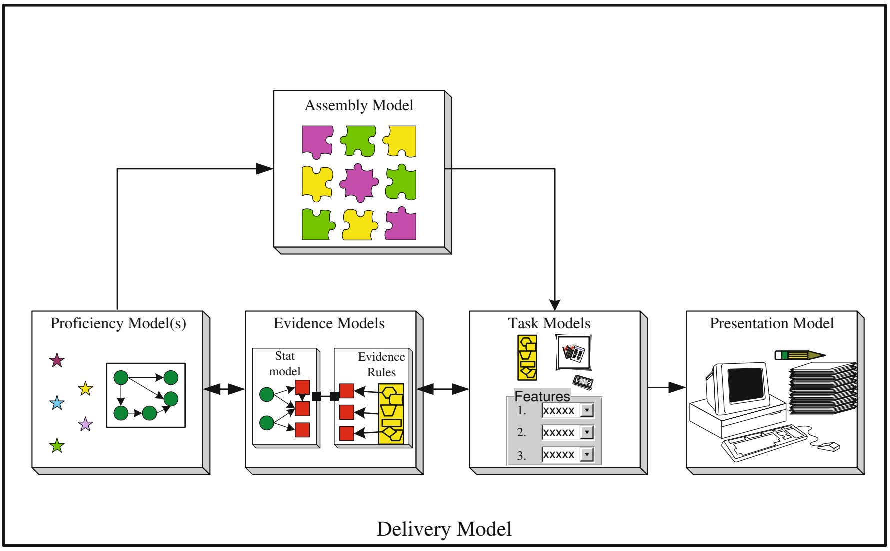

Fig. 2.1 The principle design objects of the conceptual assessment framework (CAF). These models are a bridge between the assessment argument and the operational activities of an assessment system. Looking at the assessment argument, they provide a formal framework for specifying the knowledge and skills to be measured, the conditions under which observations will be made, and the nature of the evidence that will be gathered to support the intended inference. Looking at the operational assessment, they describe the requirements for the processes in the assessment delivery system.

Reprinted from Mislevy et al. (2004) with permission from the Taylor & Francis Group.

What Are We Measuring? *The Proficiency Model*

A *proficiency model* defines one or more variables related to the knowledge, skills, and abilities we wish to measure. A simple proficiency model characterizes a student in terms of the proportion of a domain of tasks the student is likely to answer correctly. A more complicated model might characterize a student in terms of degree or nature of knowledge of several kinds, each of which may be required in different combinations in different tasks. It may address aspects of knowledge such as strategy use or propensity to solve problems with certain characteristics in certain situations. Looking ahead, the proficiency model variables will be the subset of the variables in a Bayesian net that accumulate evidence across tasks.

A closer look at the proficiency model in Fig. 2.1 reveals two kinds of elements. On the right is a graphical structure, a representation of the kinds of statistical models that are the focus of this book. On the left are a number of stars that represent claims. Claims are what users of assessments want to be able to say about examinees, and are the basis of score reports. A reporting rule maps information from probability distributions for proficiency model variables to summary statements about the evidence a student's performance provides it to support a claim.

Example 2.2 (Calculus Proficiency Model; Example 2.1 Continued). *Given that the primary purpose of the assessment is placement, only one variable is necessary in the proficiency model. This is a discrete variable whose levels correspond to the various placement options:* `Remedial Class`, `1st Semester Freshman`, `2nd Semester Freshman`, `1st Semester Sophomore`, `2nd Semester Sophomore`, `Junior Math Classes`. *Fig. 2.2 shows the graphical representation of this model. If there were a secondary purpose of trying to diagnose problems in low performing students, there might be a need for additional proficiency variables that would accumulate evidence about more specific skills. However, in a short test, the designers typically need to choose between good reliability for the main variables and good differential diagnosis for problems in the assessment. University C could use two tests: This placement test first, followed by a diagnostic test just for students placed into the remedial class, addressing only claims concerning precalculus skills and accumulating evidence at a grainsize that matches the instructional modules.*

Proficiency Level

Fig. 2.2 The proficiency model for a single variable, *Proficiency Level*. The *rounded rectangle* with the *circle symbol* represents the proficiency variable, and the *square box* with the table represents its probability distribution

Reprinted with permission from ETS.

Associated with each level of the proficiency variable are one or more claims. Which claim is associated with which level depends on how the various skills are taught in the calculus series. For example, the level `2nd Semester Freshman` *would be associated with all of the claims that constitute the kinds of performances in the kinds of tasks we would want a student successfully completing that course to be able to do. If multivariate calculus is not taught*

Springer

http://www.springer.com/978-1-4939-2124-9

Bayesian Networks in Educational Assessment
Almond, R.G.; Mislevy, R.J.; Steinberg, L.; Yan, D.;
Williamson, D.
2015, XXXIII, 662 p. 155 illus., 87 illus. in color.,
Hardcover
ISBN: 978-1-4939-2124-9