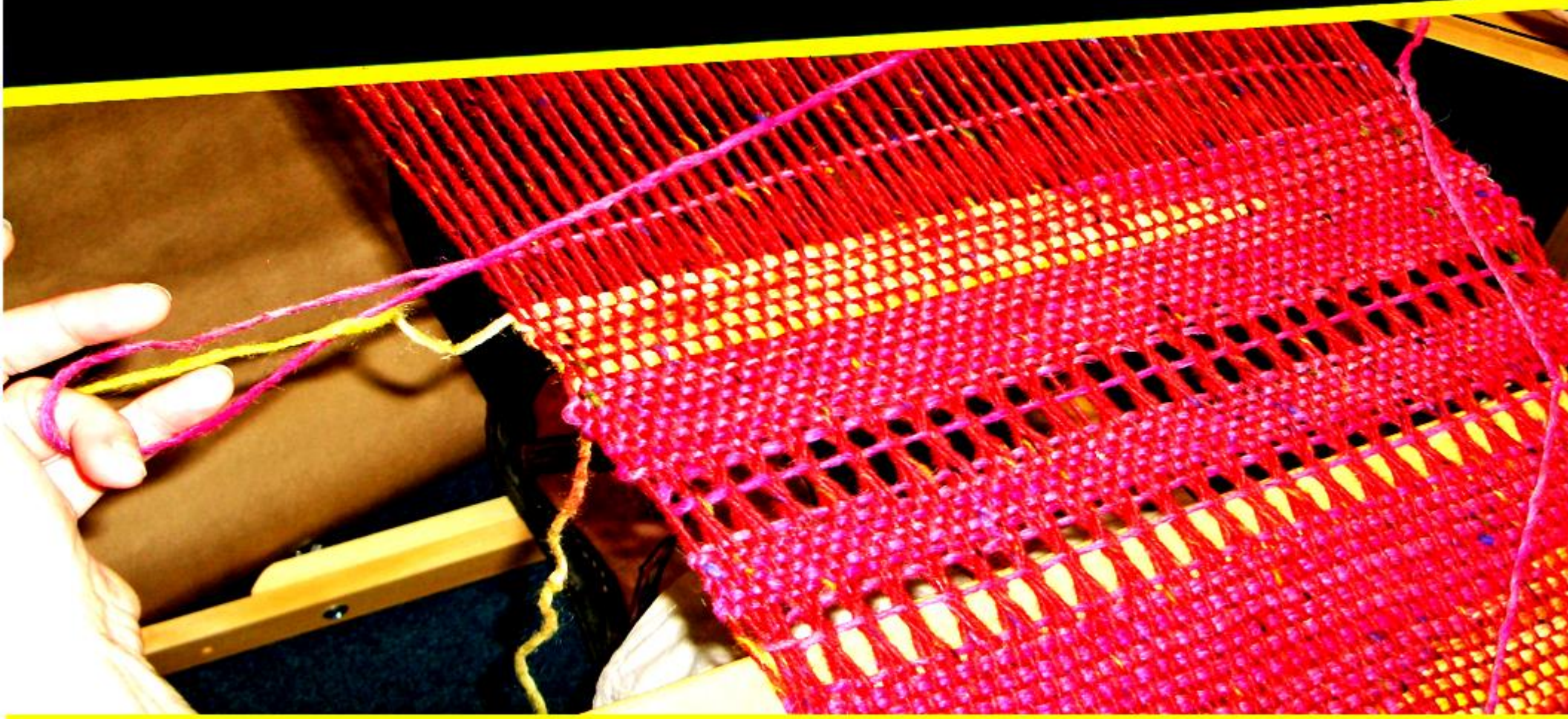
Evidence-centered design (ECD) is explained and illustrated in this working example.

ECD FOR DUMMIES

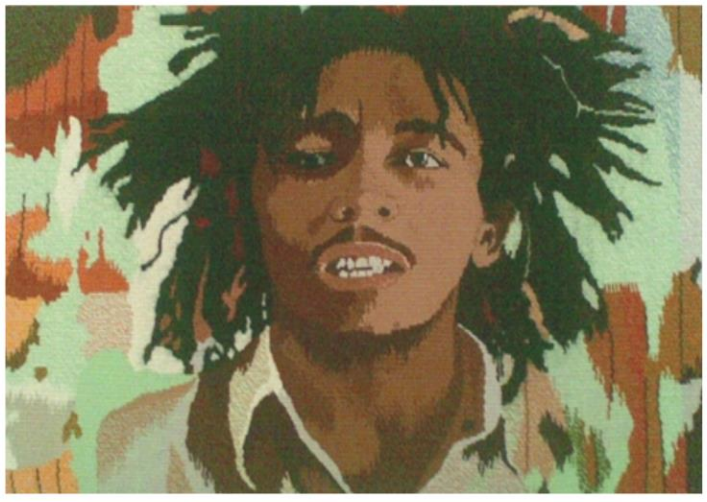


Valerie Shute, Yoon Jeon Kim, and Rim Razzouk

Table of Contents

A Metaphor

Weaving is a process of interlacing threads of different colors and fibers (e.g., cotton, silk, and wool). The goal is to produce a beautiful tapestry, rug, or fabric. It requires very careful design at the outset of the process, with each thread having an appropriate time and place.

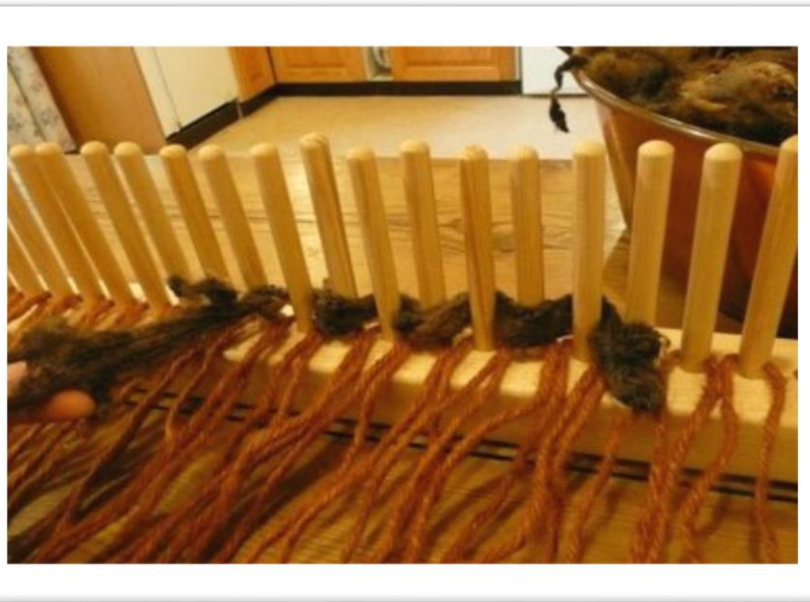


A Metaphor

ECD is similarly a design process that produces beautiful (valid and reliable) assessments of various constructs relating to knowledge, skills, values, feelings, and beliefs. These constructs are *unobservable* and thus theoretical.

"Threads" of evidence

ECD's strength comes from carefully identifying and weaving together relevant evidence to inform the construct. Evidence (which is *observable* and thus empirical) can be quantitative or qualitative, strong or weak, and relate to one or more constructs.

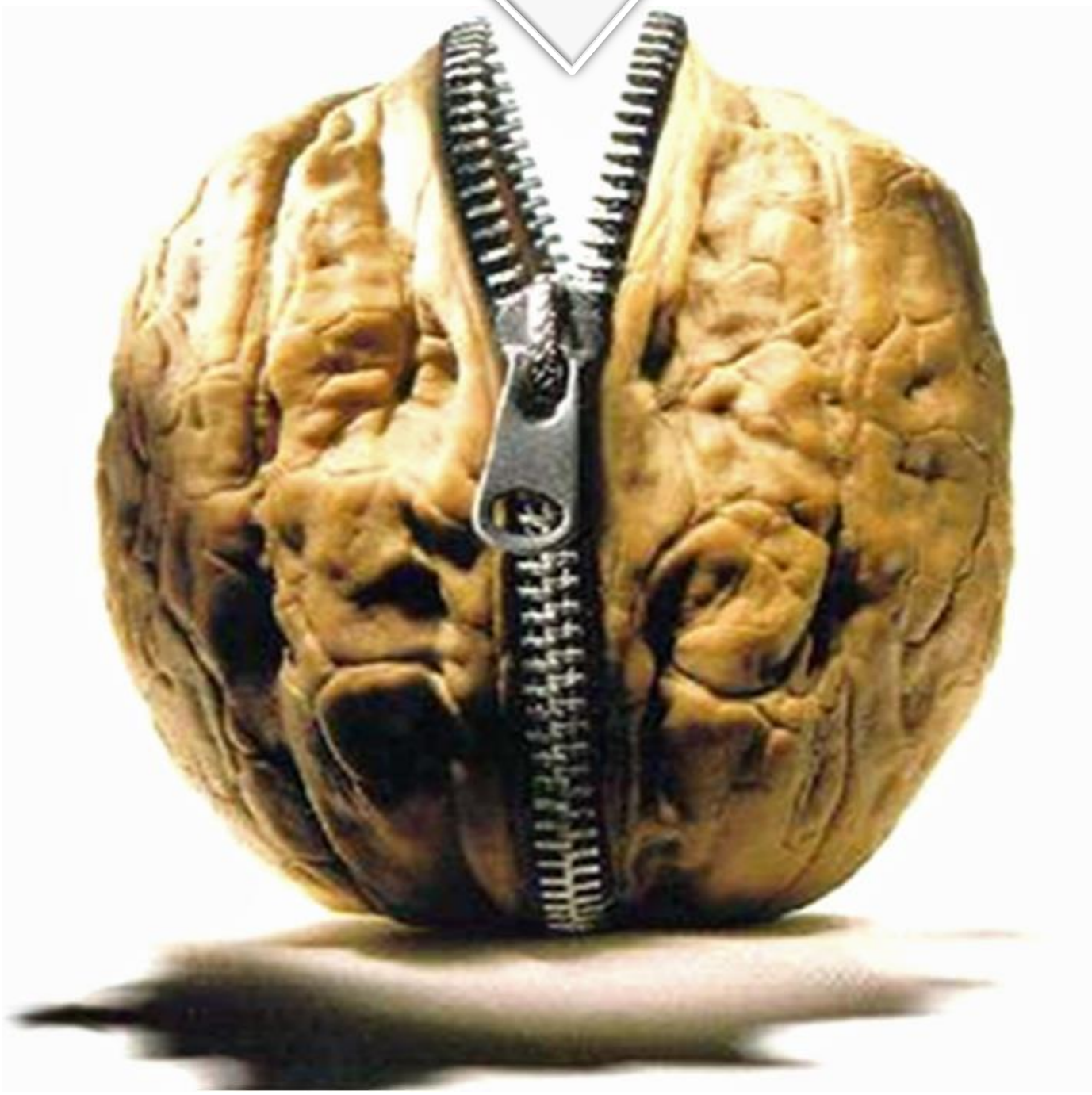


The process of weaving evidence

You can't directly observe Jay's jealousy or Ché's chess knowledge, but you <u>can</u> observe relevant behaviors and make inferences about those attributes.

TOC

ECD is a systematic way to design assessments. It focuses on the evidence (performances and other products) of competencies as the basis for constructing excellent assessment tasks.



TOC

Why Use ECD?

The world has changed a lot in the past 100 years. Education has not.

Classroom photo, 1910.

Classroom photo, 2010.

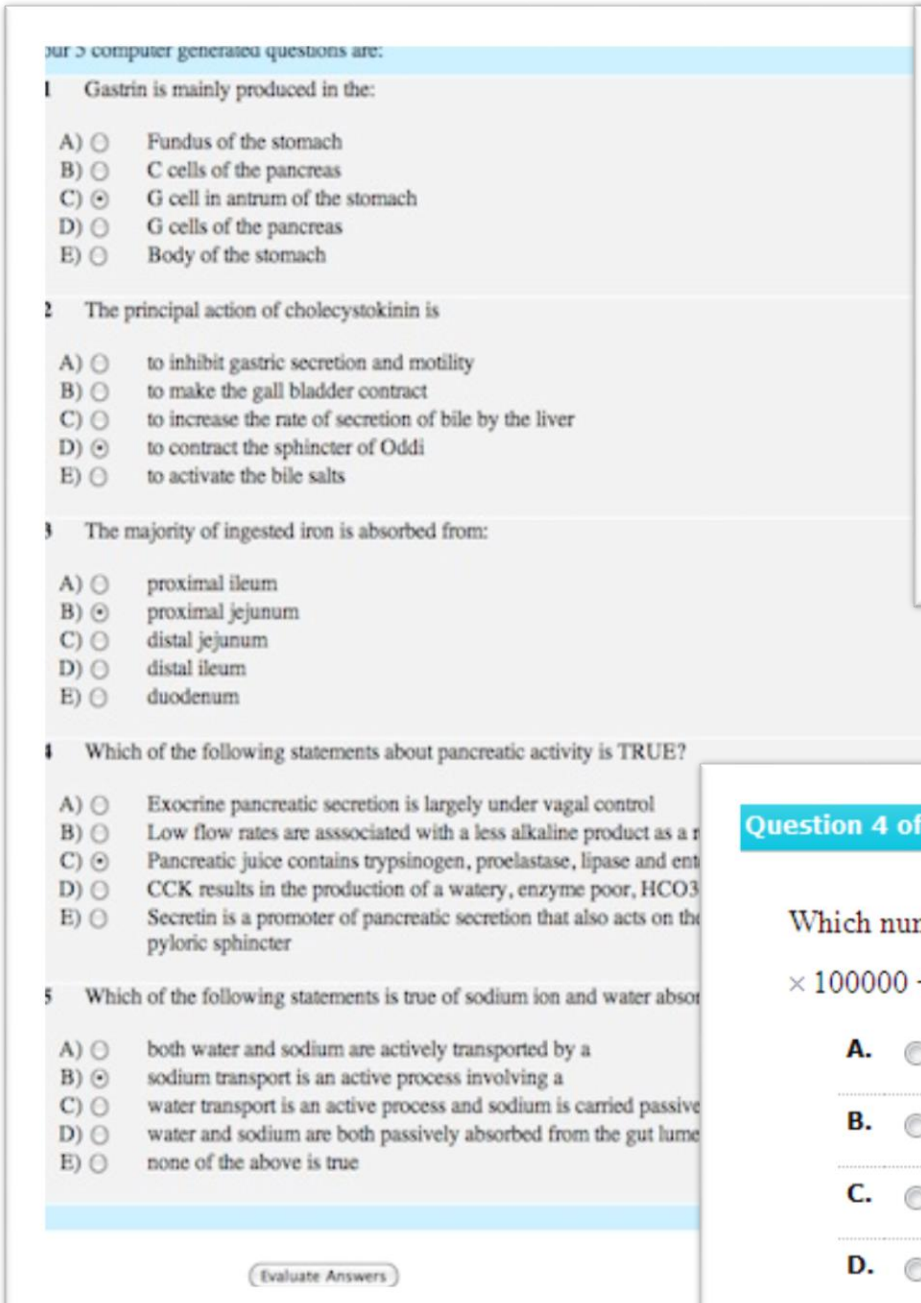
The demands associated with living in a highly technological and globally competitive world require today's students to develop a very different set of skills than their parents (and grandparents) needed.

In the past, a person who acquired basic reading, writing, and math skills was considered to be sufficiently literate. But when faced with highly technical and complex problems, the ability to think creatively, critically, collaboratively, systemically, and then communicate effectively is essential. These are examples of what many are calling *21st century competencies*.

TOC

Why Use ECD?

So education needs to change, and we also need a new approach to assessment because (a) succeeding in today's complex, dynamic world is not easily or optimally measured by multiple-choice responses on simple knowledge tests, and (b) typical multiple-choice tests are too narrow, superficial, and don't support either deep learning or the acquisition of complex competencies.



pur 5 computer generated questions are:

1 Gastrin is mainly produced in the:

A) ○ Fundus of the stomach
B) ○ C cells of the pancreas
C) ⊙ G cell in antrum of the stomach
D) ○ G cells of the pancreas
E) ○ Body of the stomach

2 The principal action of cholecystokinin is

A) ○ to inhibit gastric secretion and motility
B) ○ to make the gall bladder contract
C) ○ to increase the rate of secretion of bile by the liver
D) ⊙ to contract the sphincter of Oddi
E) ○ to activate the bile salts

3 The majority of ingested iron is absorbed from:

A) ○ proximal ileum
B) ⊙ proximal jejunum
C) ○ distal jejunum
D) ○ distal ileum
E) ○ duodenum

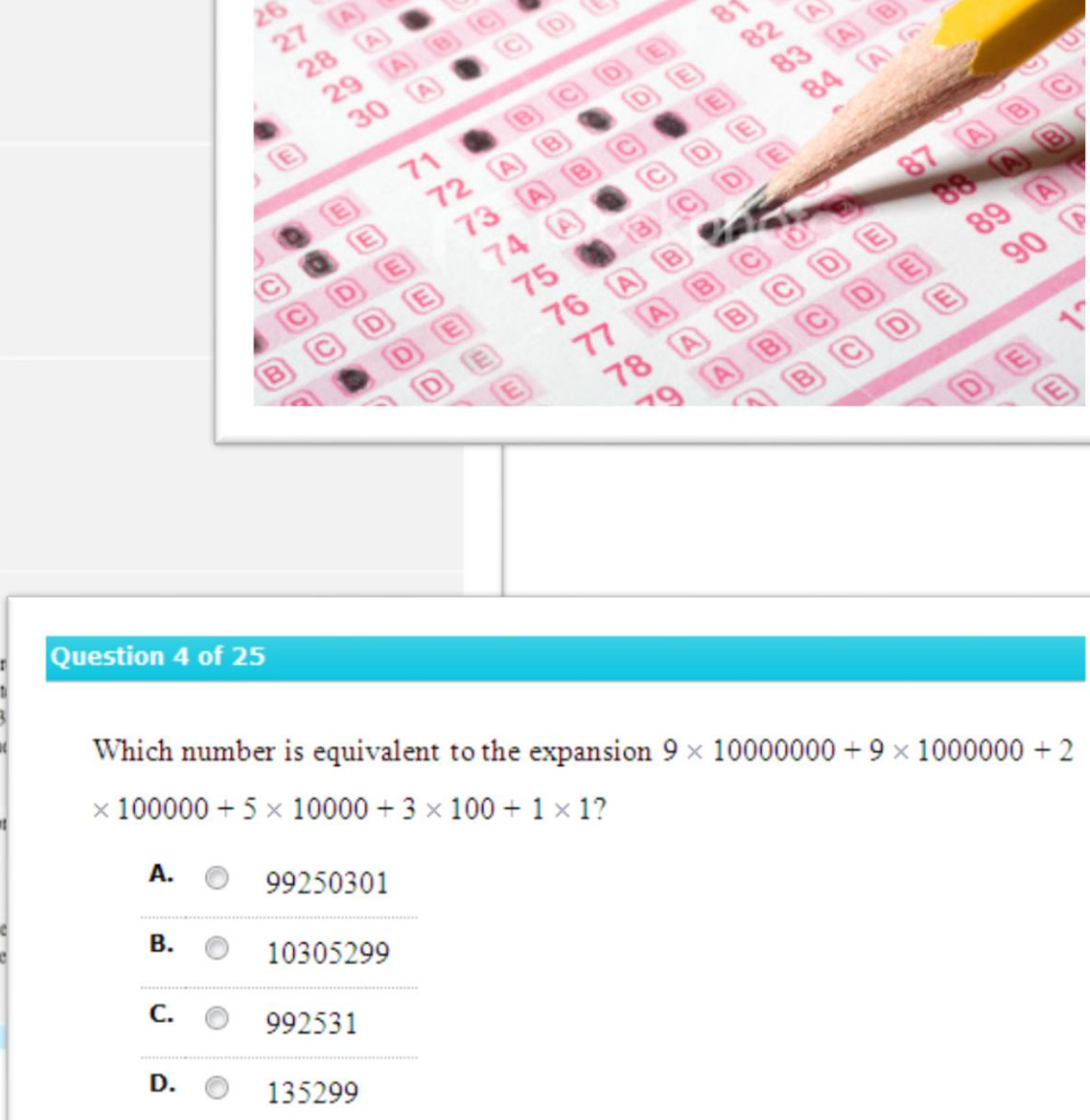
4 Which of the following statements about pancreatic activity is TRUE?

A) ○ Exocrine pancreatic secretion is largely under vagal control
B) ○ Low flow rates are asssociated with a less alkaline product as a r
C) ⊙ Pancreatic juice contains trypsinogen, proelastase, lipase and ent
D) ○ CCK results in the production of a watery, enzyme poor, HCO3
E) ○ Secretin is a promoter of pancreatic secretion that also acts on th pyloric sphincter

5 Which of the following statements is true of sodium ion and water absor

A) ○ both water and sodium are actively transported by a
B) ⊙ sodium transport is an active process involving a
C) ○ water transport is an active process and sodium is carried passive
D) ○ water and sodium are both passively absorbed from the gut lume
E) ○ none of the above is true

Evaluate Answers



Question 4 of 25

Which number is equivalent to the expansion $9 \times 10000000 + 9 \times 1000000 + 2 \times 100000 + 5 \times 10000 + 3 \times 100 + 1 \times 1$?

A. ○ 99250301

B. ○ 10305299

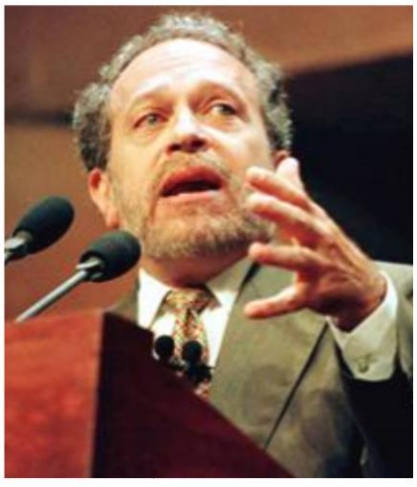
C. ○ 992531

D. ○ 135299

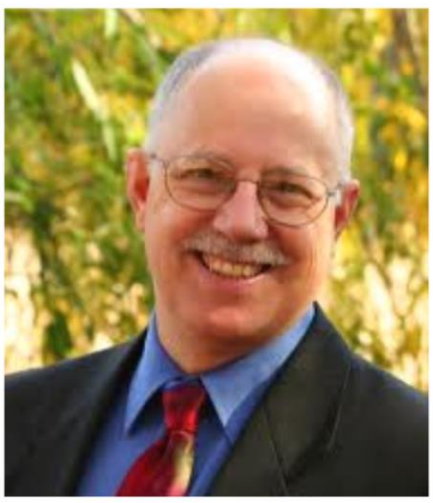
TOC

Why Use ECD?

Standardized tests are monstrously unfair to many kids. We're creating a one-size-fits-all system that needlessly brands many young people as failures, when they might thrive if offered a different education where progress was measured differently.



Robert Reich



Robert Mislevy

ECD *provides 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.*

TOC

Why Use ECD?

Any assessment collects information about a person that lets you make inferences about his or her competencies and other attributes. Accurate inferences support smart decisions that can promote learning. ECD provides an approach that yields accurate inferences. It also moves us toward the right-side column of the table below (adapted from the National Research Council, 1996).

Less Focus on Assessing	More Focus on Assessing
Learning outcomes	Learning processes
What is easily measured	What is most highly valued
Discrete, declarative knowledge	Rich, authentic knowledge and skills
Content knowledge	Understanding and reasoning, within and across content areas
What learners do not know	What learners understand and can do
By teachers	By learners engaged in ongoing assessment of their work and that of others

TOC

Claim

ECD should be used as the framework for new assessments because it:

- Can yield *valid assessments* for different purposes (e.g., formative assessments to support learning, summative exams).
- Provides for *accurate estimates* of complex competencies, dynamic performances, and other hard-to-capture-and-analyze data.
- Can *aggregate information* from various sources (such as qualitative and quantitative data and *in situ* learning).
- Affords *transparency* to stakeholders (and thus accountability) via evidentiary reasoning to support claims.

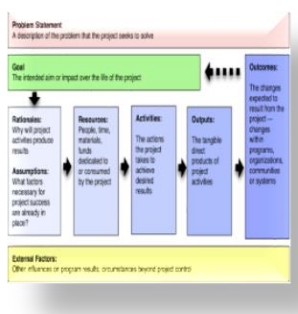
TOC

Very Brief History of ECD

- ECD originated at Educational Testing Service in 1997 out of the minds of *Robert Mislevy*, *Linda Steinberg*, and *Russell Almond*. It is a principled framework for designing, developing, and delivering valid assessments.

- ECD builds on the vision of Samuel Messick, "*The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics*."

- In ECD, all of the various parts and processes of an assessment get their meaning from an *assessment argument* (i.e., a series of statements where the final statement is a conclusion or claim which follows logically from the preceding statements or premises).



For more on ECD from one of its founders see: **http://ecd.ralmond.net/ecdwiki/ECD/**

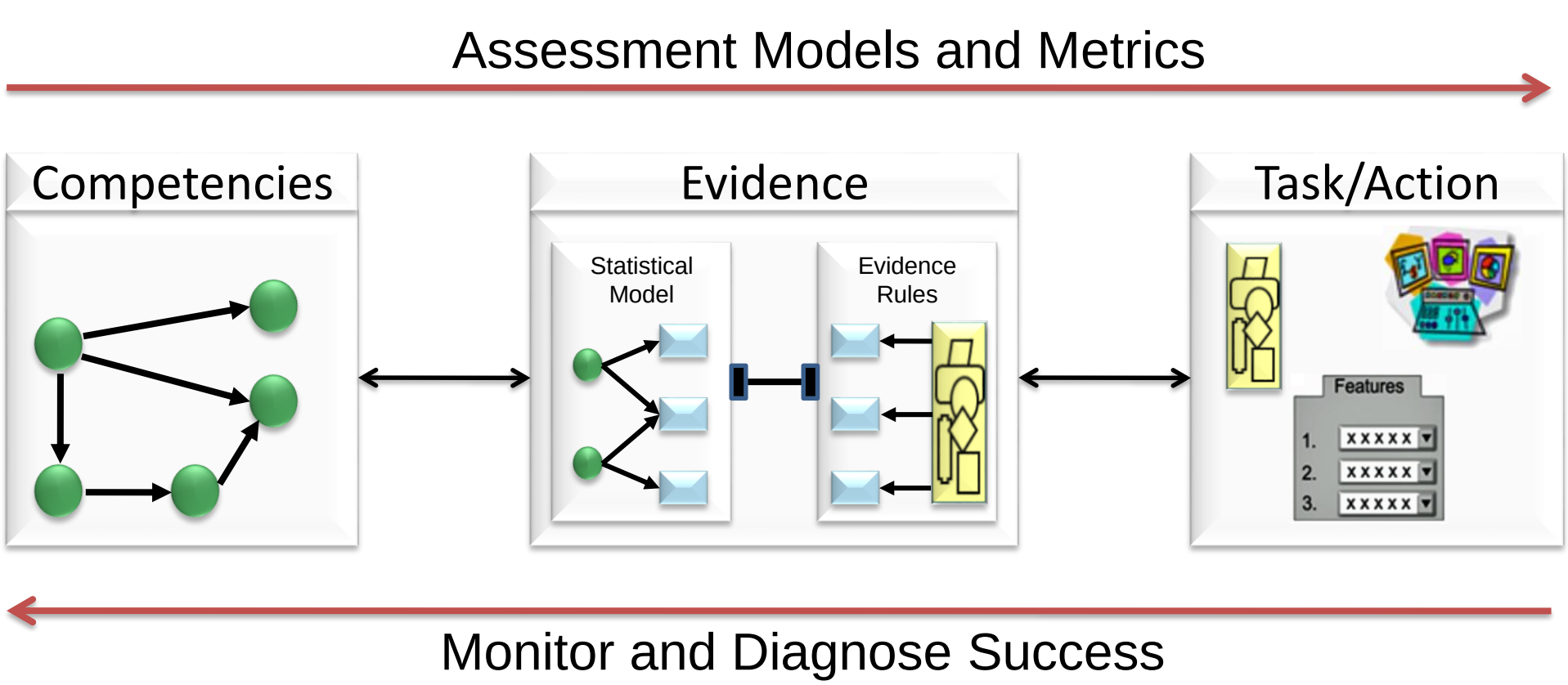
TOC

What is ECD?

- ECD has two main functions. It provides a way to reason about assessment design, and a way to reason about a person's performance (diagnostically speaking).
- ECD can be used to design assessments of all kinds, and is especially suited for assessments that involve complex competency models and dynamic, interactive environments that lie beyond the analytic capabilities of simpler assessments.

What is ECD?

- ECD, in its simplest form, can be described by three main models:
  - Competency Model
  - Evidence Model
  - Task (or Action) Model
- Below is a picture showing the flow between the models. We'll go through each of the models in turn.

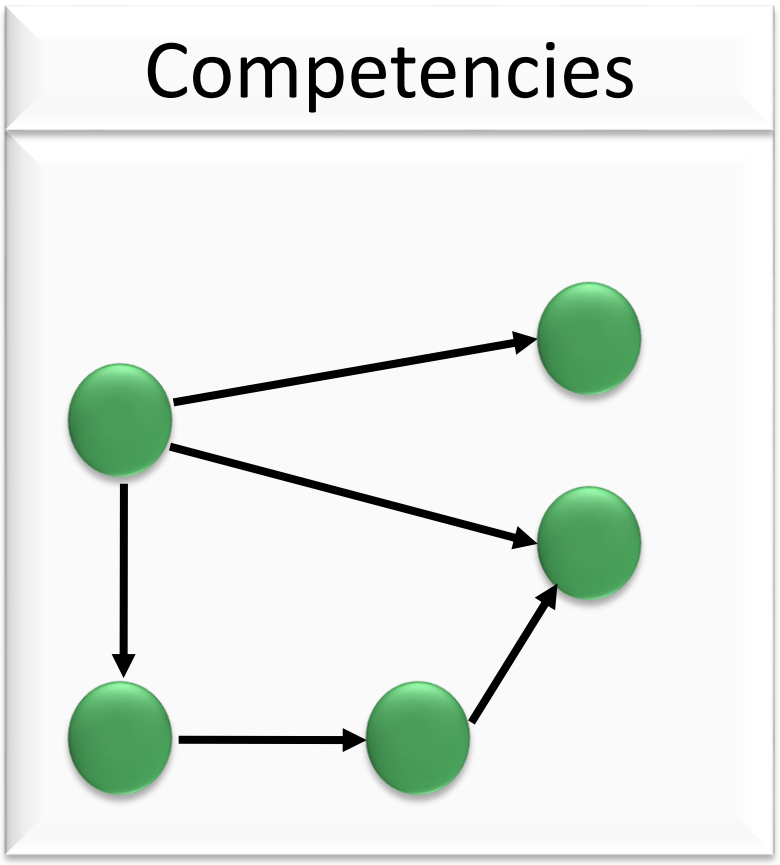


The red arrow heading left-to-right shows reasoning about assessment design (competency to evidence to task model). And the arrow going from right-to-left demonstrates reasoning about a person's performance.

TOC

Competency Model (CM)

What collection of knowledge, skills, and other attributes should be assessed?



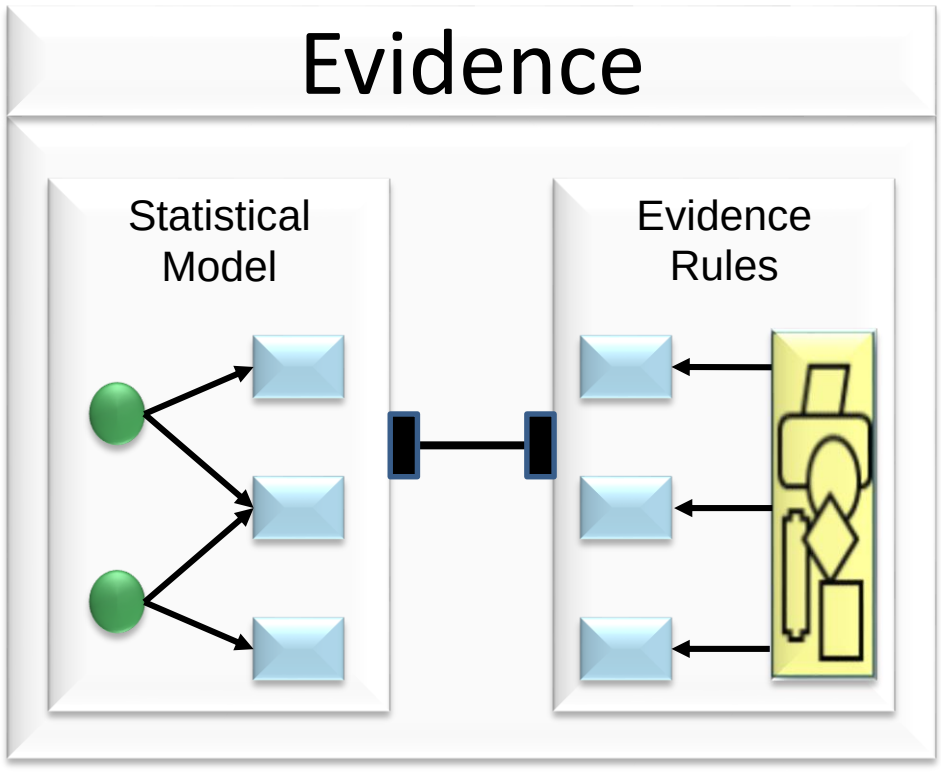
- Variables (green circles) in the CM describe knowledge, skills, and other attributes about which inferences are intended.
- Inferences can be at various grain sizes, from general (e.g. *Maya's math skills are high*) to more specific (*Jeb is having serious problems solving linear equations*).

- The term "student model" may be used to refer to a student instantiated version of the CM—like a profile. Values in the student model express current beliefs about a learner's level on each variable within the CM.

TOC

Evidence Model (EM)

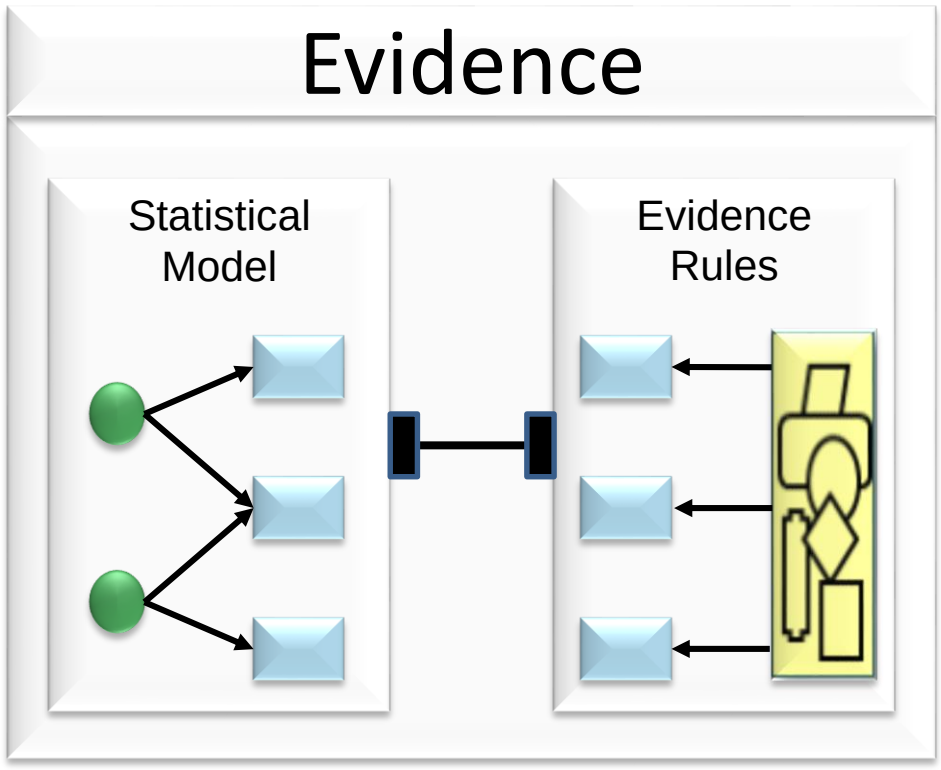
What behaviors should reveal different levels of the targeted competencies?



- The EM analyzes a person's interactions with, and responses to a given problem. This is the *evidence* which informs the CM variables.
- The EM consists of two parts: (a) Evidence Rules and (b) Statistical Model.
- *Evidence Rules* (i.e., rubrics or scoring model) take as input the work product (shown as the yellow rectangle) that comes from the person's interaction with a task or learning environment. Depending on the type of task, the work product might be a short answer, a piece of artwork, a sequence of actions, and so on. As output, evidence rules produce observable variables (i.e., scores, shown by the blue boxes) that are evaluative summaries of the work products.

Evidence Model (EM)

What behaviors should reveal different levels of the targeted competencies?



- The *Statistical Model* expresses the relationship, in probability or logic, between the CM variables and the observable variables (scores). It enables updating the CM variables in a way that combines scores across tasks or performances.

- The *Statistical Model* may be as simple as number-right scoring for a single competency variable, or it may use Bayes net software to update competency variables with conditional probabilities.

- Basically, a conditional probability gives an estimate for the likelihood that *Person X* is at a certain level of proficiency for *Skill Y* given all relevant data collected so far.

TOC

Task Model (TM)

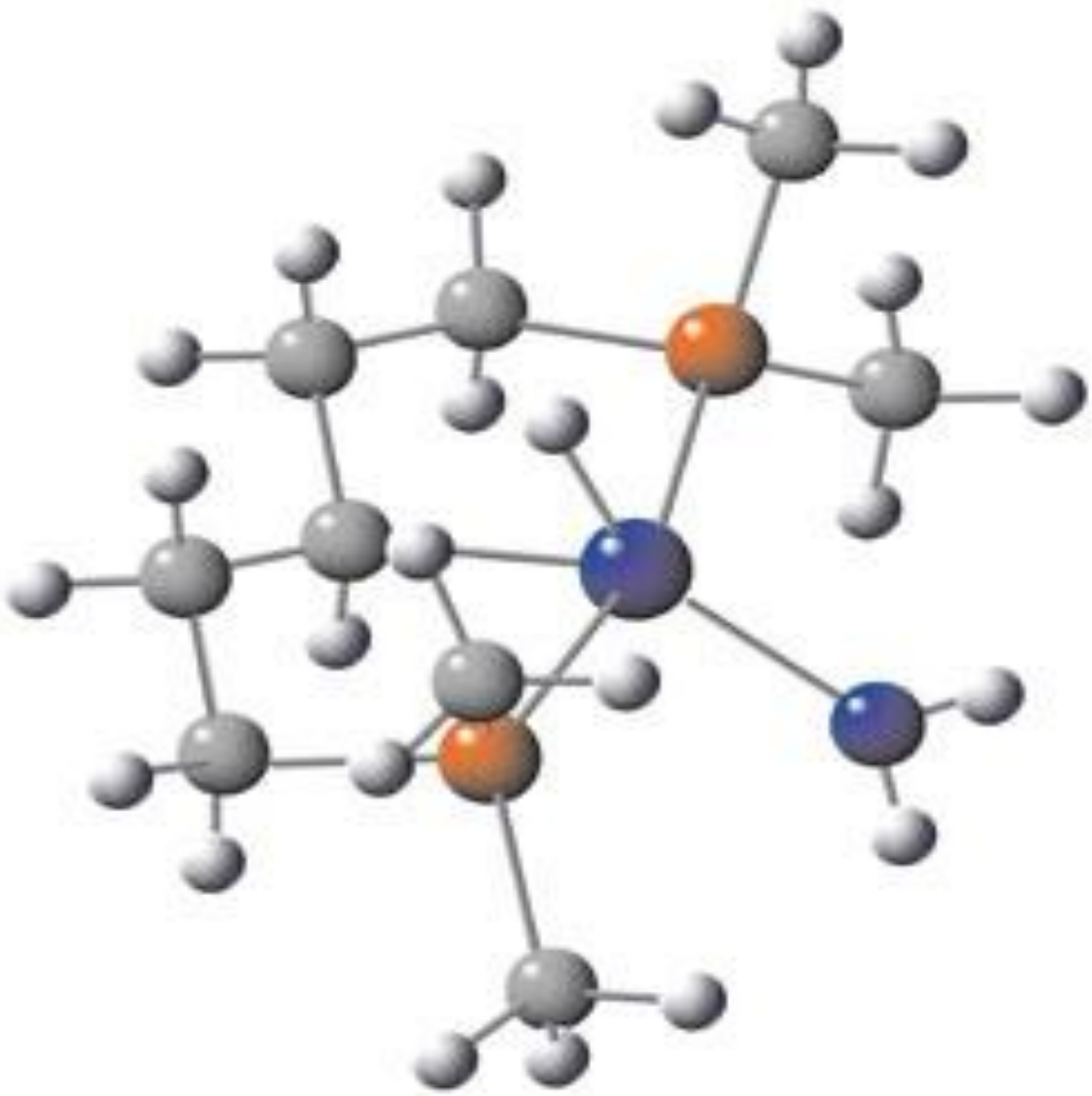
What tasks/situations can elicit the behaviors that make up the evidence?

- The TM provides a framework for describing and constructing *situations* with which a person will interact to provide evidence about aspects of competencies.
- Situations are described in terms of: (a) *presentation* format (e.g., on the computer or tennis court), (b) specific *work product* (e.g., haiku or geometry proof), and (c) other variables (e.g., difficulty level).

- When ECD assessments are used within games, we use the term *action model* instead of task model. This reflects the fact that we are dynamically modeling learners' action sequences which form the basis for drawing evidence and inferences. The action model in a gaming situation defines the sequence of actions, and each action's indicators of success. Actions represent the things that learners do to complete the mission or solve a problem.

TOC

Applying ECD

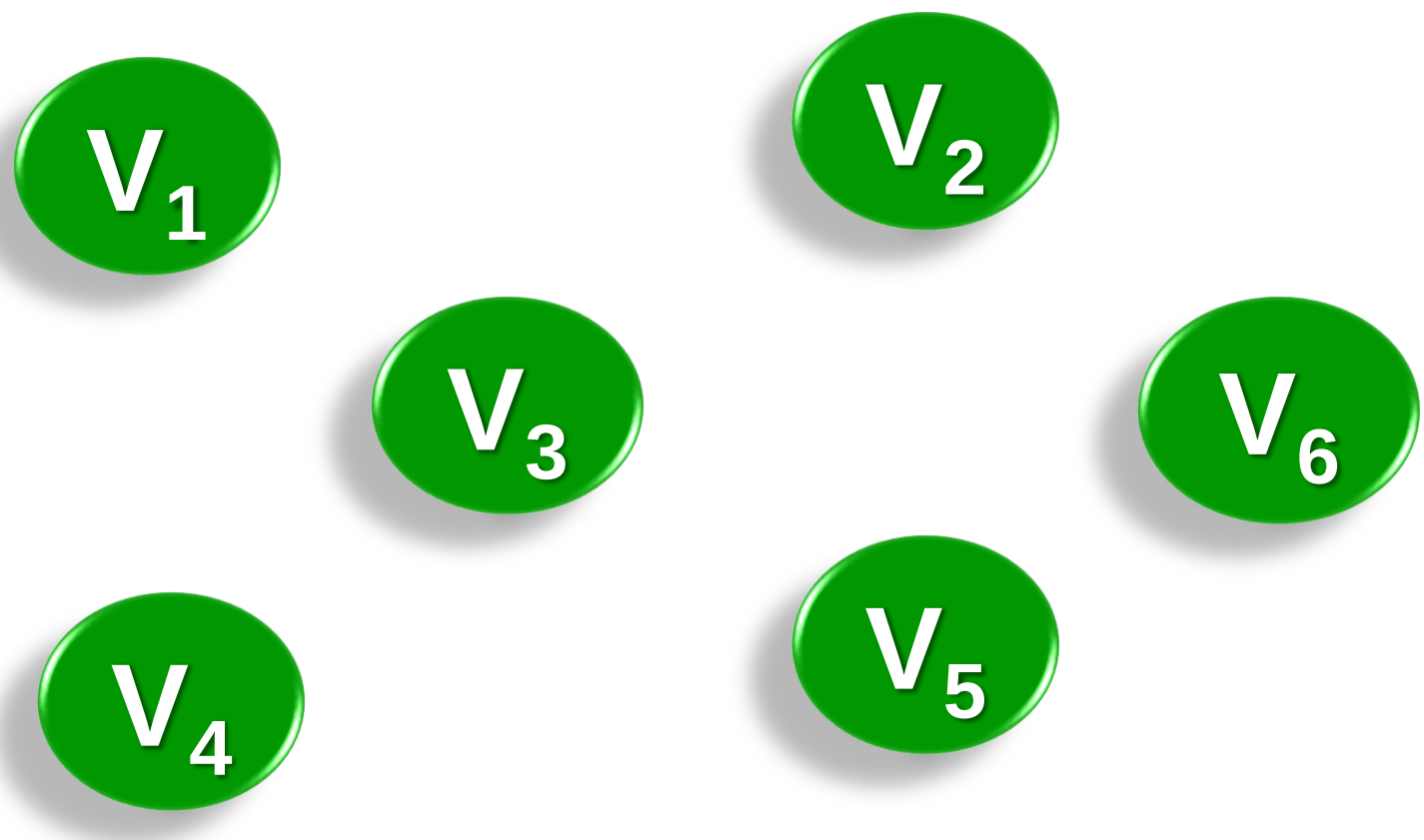


- Just like any scientist building a model, you begin your ECD-based assessment by specifying the variables of interest, along with a structure of the variables for your competency model.

- The structure of the variables is usually explained by what's called a probability distribution. We'll see examples of that in a minute.

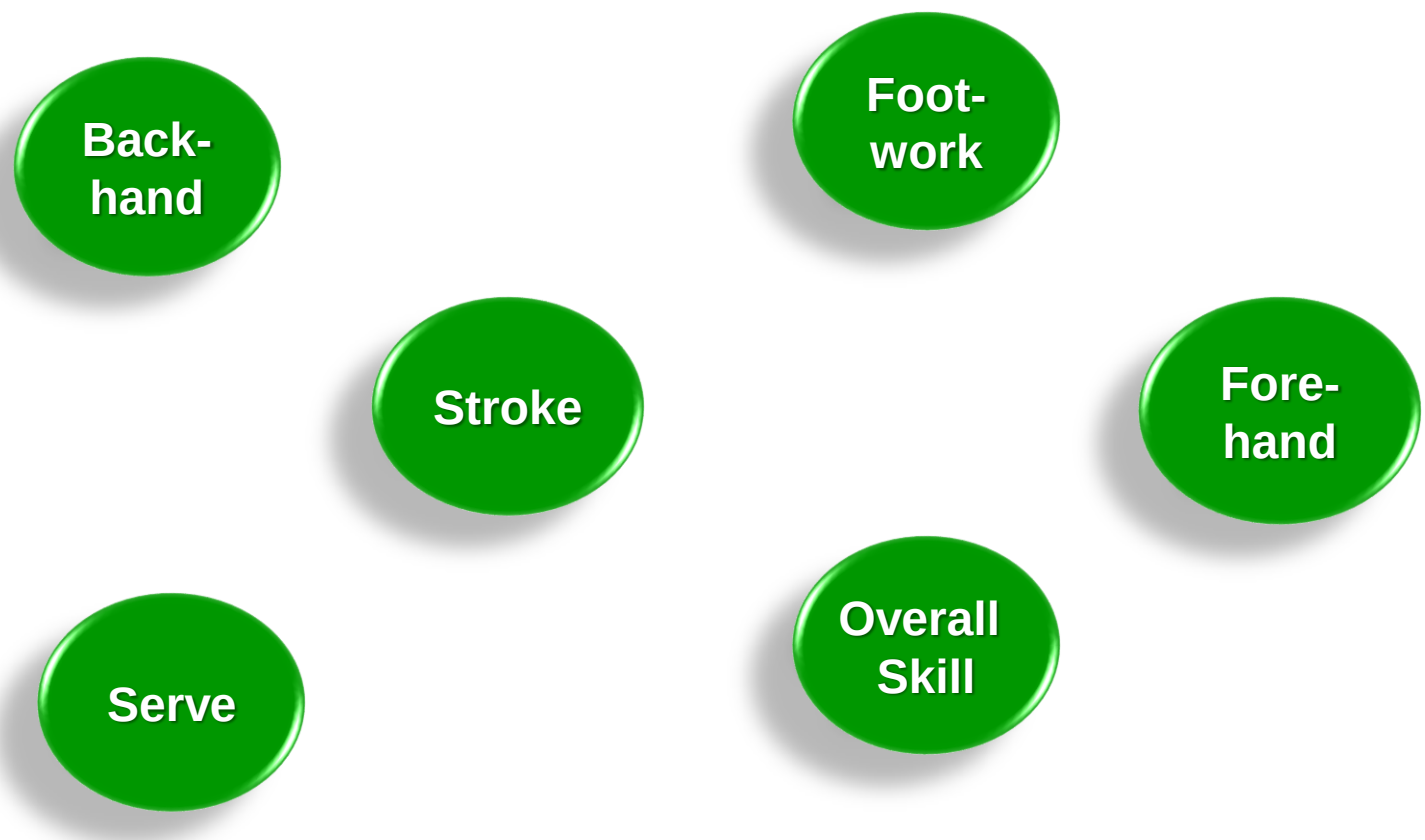
TOC

Selecting CM Variables



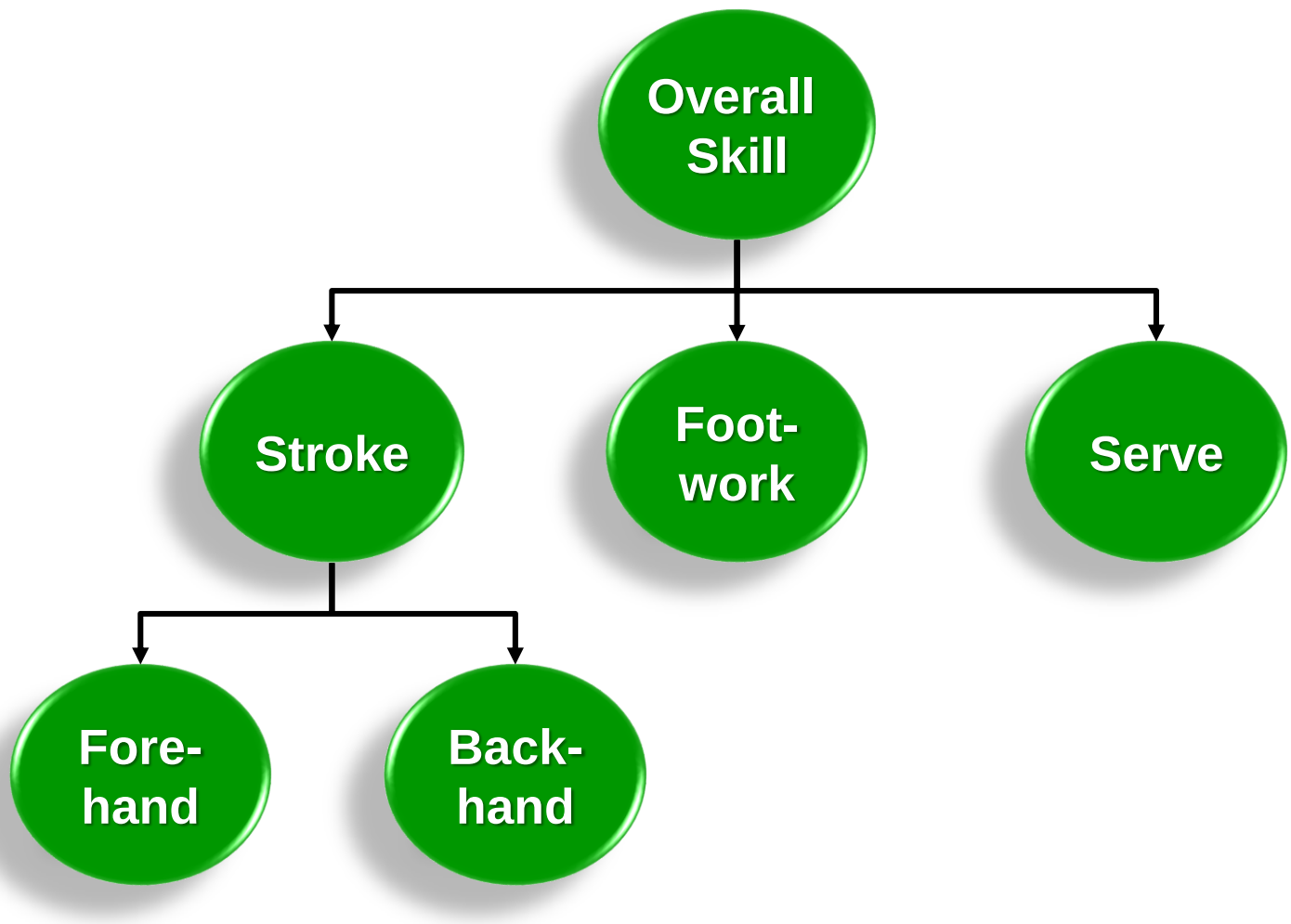
- So, what do you want to measure? Remember, the CM is a collection of *variables* that correspond to learners' attributes such as skills, knowledge, and abilities about which you want to make claims.
- Suppose you wanted to make a claim about a person's ability to *play tennis*. What would you use for your variables? What skills does a good tennis player need to have? What do novices do?
- For this tennis-playing example, your CM will include variables such as stroke (both forehand and backhand), footwork, and serve.

Selecting CM Variables



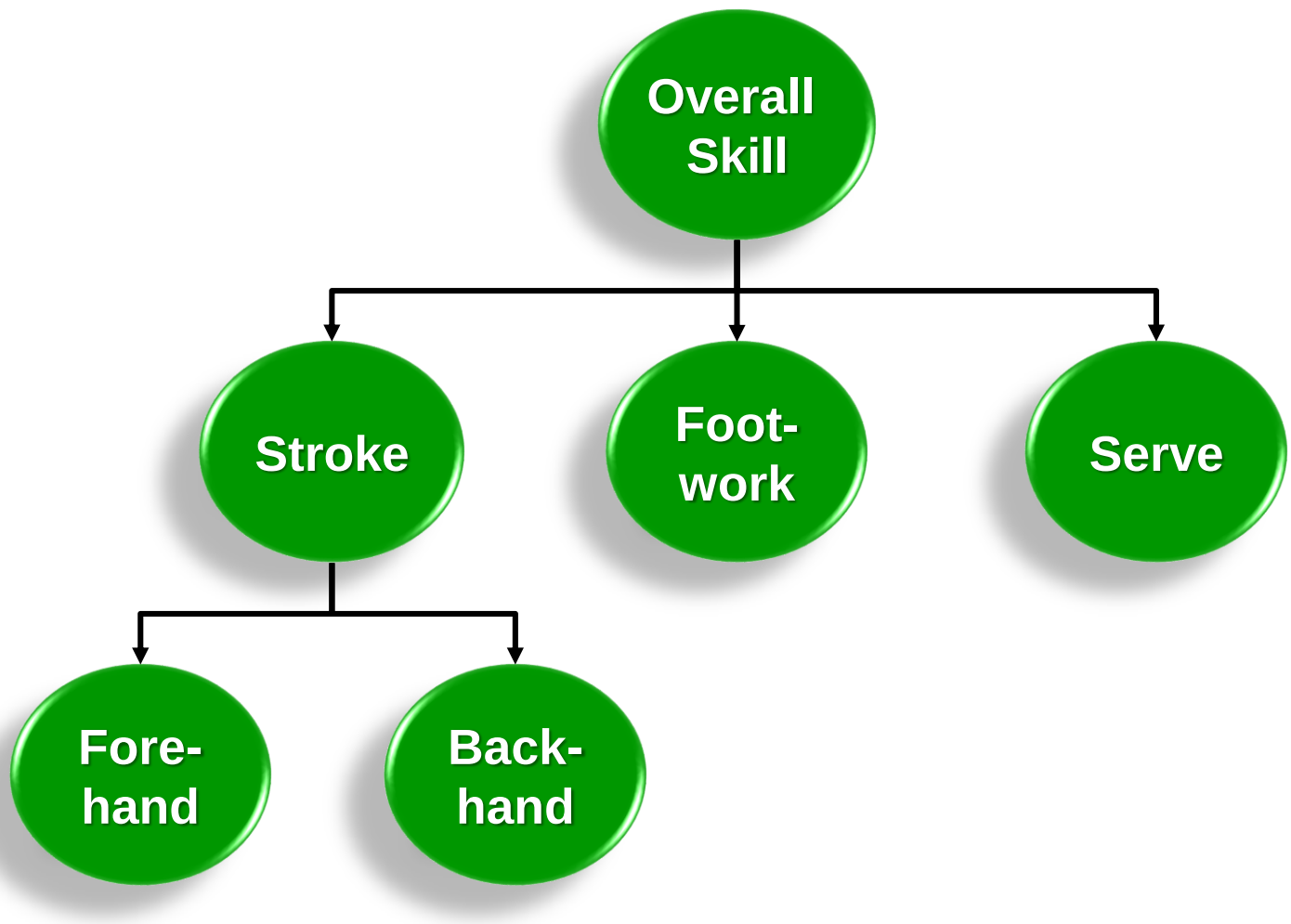
- We've come up with the variables (above) and we'll pretend like this represents the complete set of variables related to tennis-playing skill.
- Please note that the variables, just scattered around like they are, don't mean anything. To have meaning, they must be *structured* to represent their interrelationships.
- The structure is shown graphically, and may be explained as a probability distribution of the variables. Let's think about an appropriate structure for these variables.

Structuring CM - Conceptual



- Here's a possible structure of the variables. Does this seem logical to you?
- Now let's think about different "weights" per variable. For example, consider the variables *stroke* and *footwork*. Are both equally important to overall tennis skill?
- Tennis experts say that the most important skill for tennis performance is not one's stroke or serving ability, but footwork because good footwork is a precondition for a good stroke. So we need to somehow indicate the larger influence of footwork in the model, compared to stroke and serve.

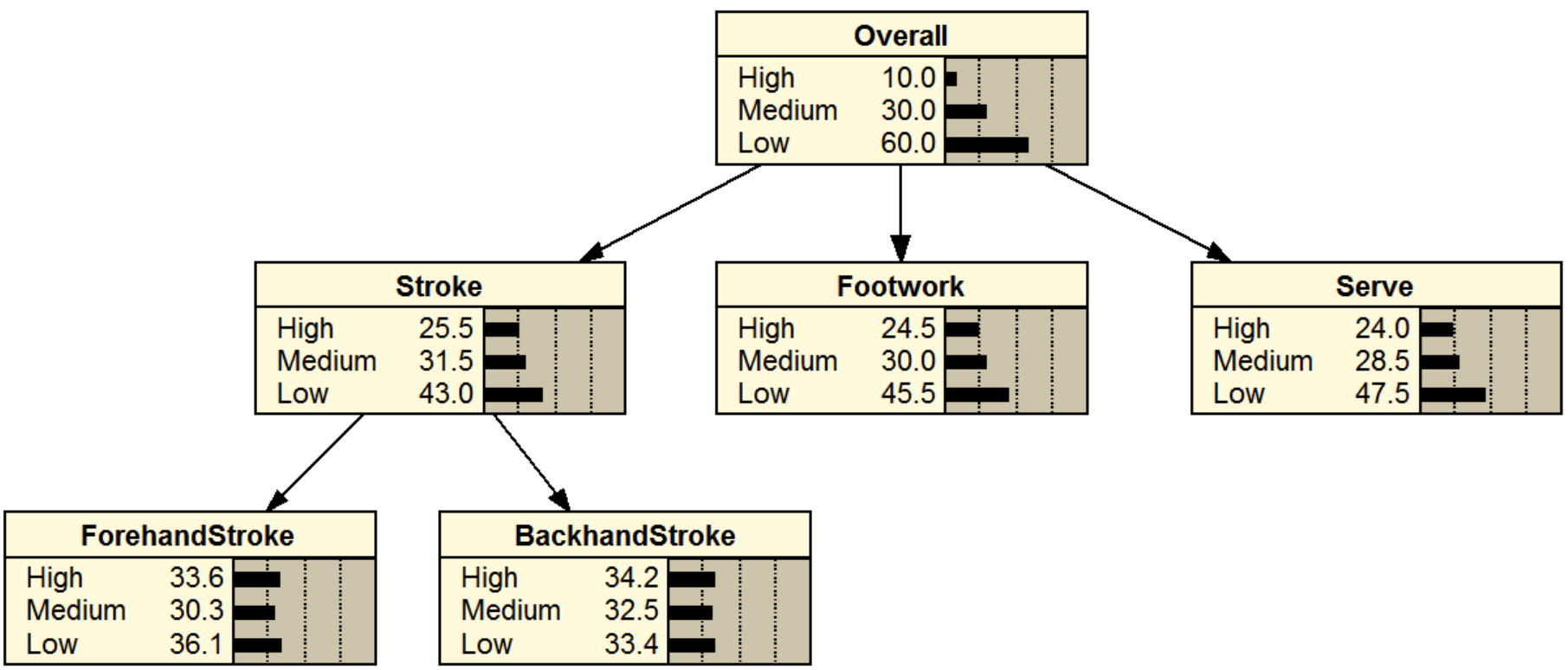
Structuring CM - Conceptual



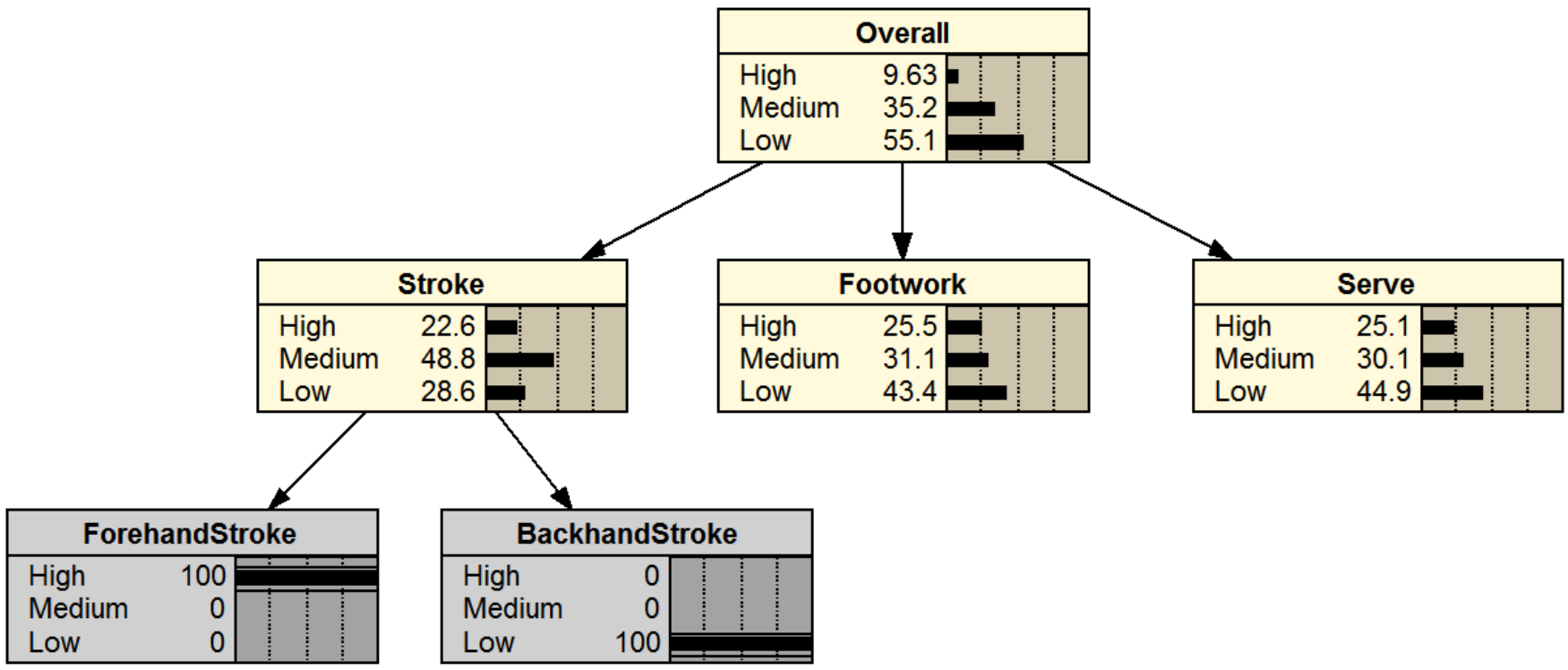
- Let’s think about another situation. An aspiring tennis player named Chaz consistently demonstrates strong and precise forehand strokes, but his backhand strokes are weak and inaccurate. While backhand strokes are usually harder to master, both are about equally important to playing tennis.
- Given this particular profile, how do you think each stroke variable (forehand vs. backhand) influences the overall stroke? Probably about medium, right? The next page shows a probability distribution illustrating the relationships.

TOC

Structuring CM - Computational

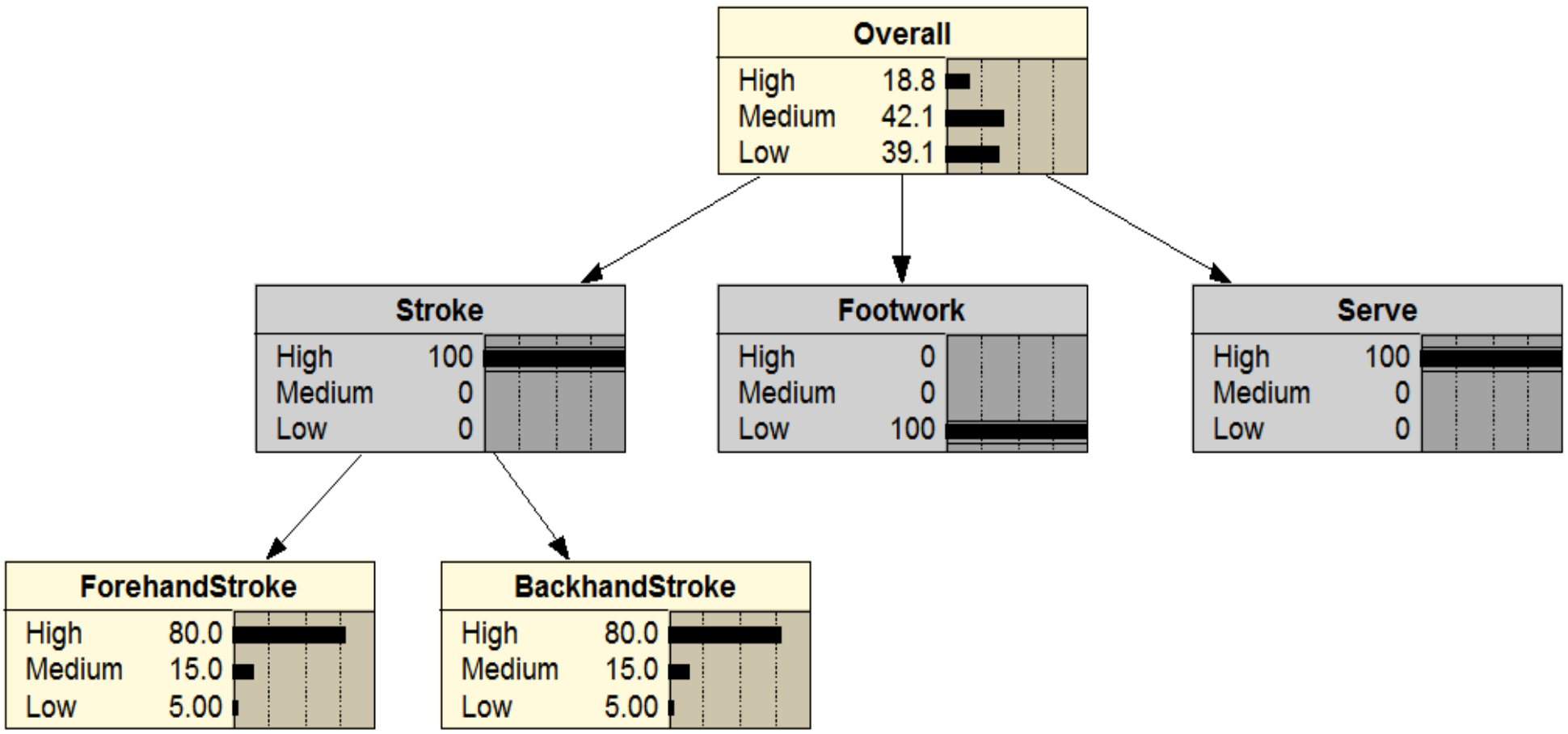


This is an example of a probability distribution of the variables.



As you probably guessed, if a player is estimated to be *high* in relation to his forehand stroke, and *low* on his backhand stroke, he's estimated as being *medium* in terms of the stroke variable.

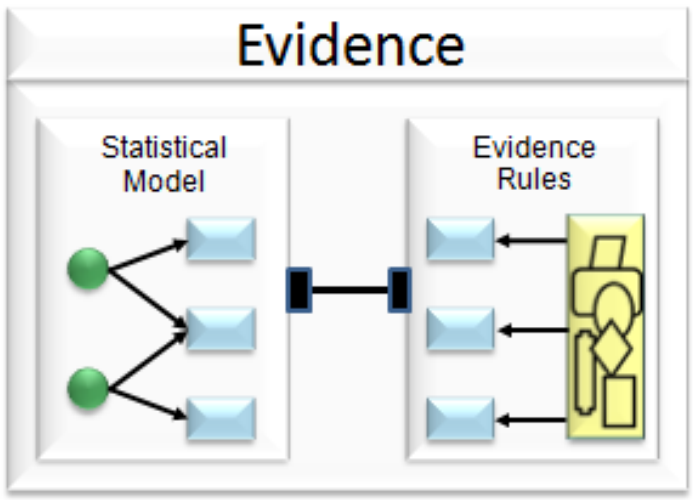
Structuring CM - Computational



- Earlier we wondered about how we could show different degrees of influence of variables on each other. To illustrate, suppose that Chaz has demonstrated a high level of skill for both the stroke and serve variables, but a poor level of footwork skill. See the network picture above.
- He's estimated as somewhere between medium and low in relation to the overall performance. That's because footwork has a relatively large influence on the overall variable, which is reflected in the probabilities.

TOC

Building the EM



- Now it's time to move our attention to building the Evidence Model. Remember—the EM determines how the observed actions can be used as evidence to update the current states of the competency model variables.
- We need to build two components of the EM: **evidence rules** and the **statistical model**.
- We'll also show how evidence rules and statistical models work together.

TOC

Building the Evidence Rules

- For our tennis example, we can create rubrics for the evidence rules. Evidence rules need to have (a) specific observations (i.e., indicators) that you want to see, and (b) information about how the observations will be scored.
- The following table illustrates scoring rules for the variable, *Forehand stroke*.

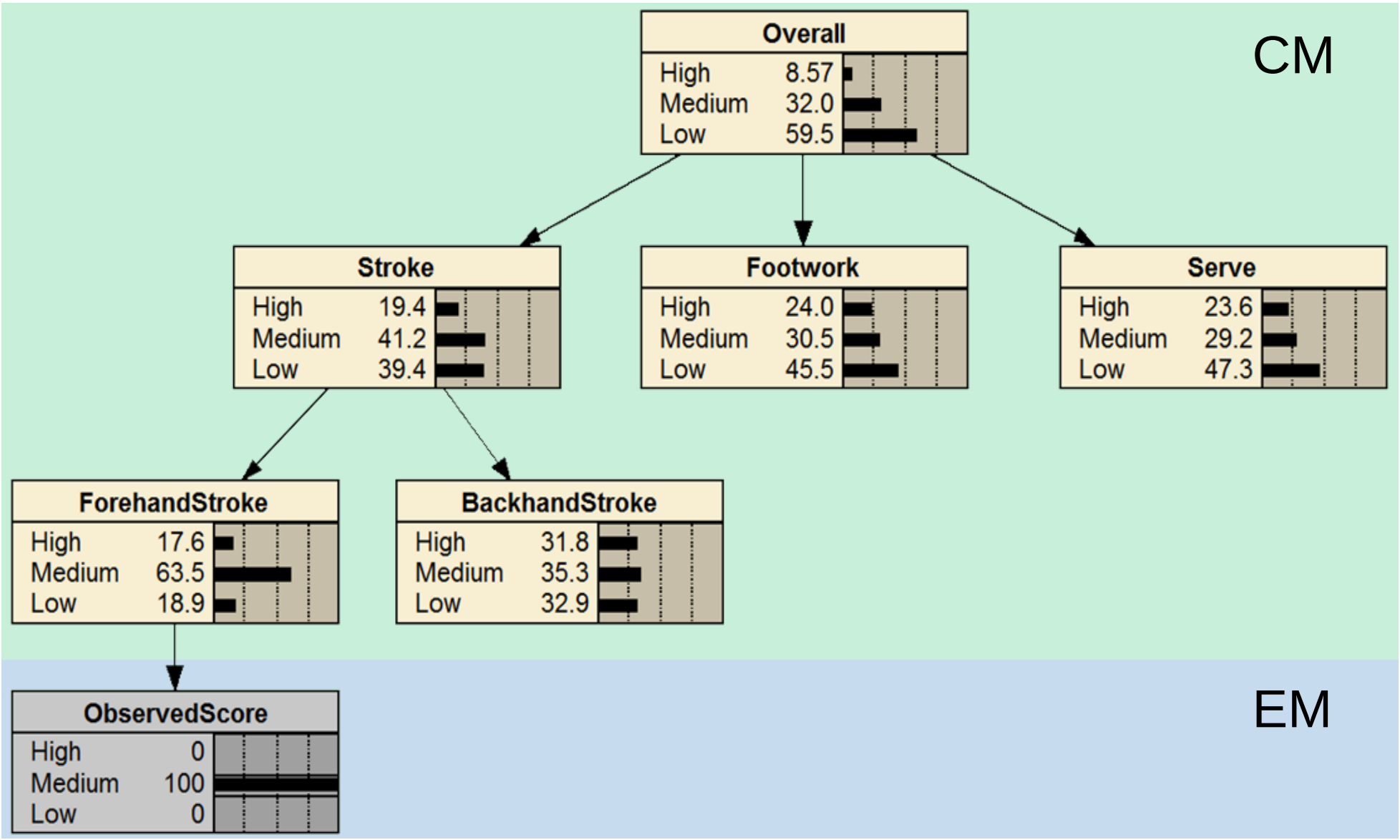
Indicators\Score	0	1	2
***Form* of forehand stroke**	Improper form	Proper form, but timing off	Proper form and good timing
***Control* of the ball's direction with forehand stroke**	The ball landed outside of the line.	The ball landed inside of the line, but in an easy spot for the opponent to hit.	The ball landed inside of the line, but in a very hard-to-hit spot for the opponent.
***Power* of stroke (longer, shorter, and follow through)**	The ball didn't cross the net, or the ball went too far.	The ball crossed the net, but provided a scoring opportunity for the opponent.	The ball crossed the net, but it was difficult for the opponent to keep the ball in play.

Building the Statistical Model

- We just specified scoring rules for our tennis example, so now we can observe a person (Li) playing tennis, and then score her forehand stroke. The highest possible score one can earn for forehand stroke is 6 (across the three indicators). Li scored a "1" on each indicator for a total score of 3. How can we use this information to estimate her current state of *forehand stroke*?
- The statistical model feeds (or statistically links) observational data into the competency model.
- First, we need to decide how to interpret the obtained data. We can use a proportion of obtained to total possible score. For instance, 3 (Li's score on forehand stroke) / 6 (the total score) = 0.50.
- Second, we can set cut-scores, as shown in the table below:
- According to the table, her score will be updated into the model as *Medium* for her forehand stroke. The next slide illustrates this updating process.

Range	States
0.68 – 1.00	High
0.34 – 0.67	Medium
0.00 – 0.33	Low

Building the Statistical Model



- As you can see, the EM statistically integrates new information into the CM, which results in an update to all CM variables.
- Based on the updated information, we can infer that Li's stroke skill is at the *medium-to-low* level, at this point in time and with just one observation. Additional observations (e.g., on her backhand stroke, footwork, and serve) will further update the model and strengthen the validity of our inferences.

TOC

Specifying the TM

- So far we have described how to build competency and evidence models. Now we need to think about *how* and *where* we will measure our targeted, important variables that make up the competency model.
- Figuring out the circumstances and settings (e.g., format and difficulty level) for tasks is the job of the task model.

Specifying the TM

- Continuing with our tennis example, what is the best environment or context that'll enable you to collect the evidence needed to estimate a person's current status with regard to his/her tennis skills?

  A. Multiple choice test
  B. 1000-word essay on tennis
  C. Let the person demonstrate his/her skills on a tennis court.

- You probably want to observe the person play tennis and evaluate the performance relative to specific indicators that are linked to variables in the competency model.

Specifying the TM

- We also need to consider specific characteristics of the physical environment where tennis play will take place. Think about players performing on three different types of court: hard court, grass, and clay. Does a player's performance vary across the different court surfaces?

- Rafael Nadal grew up playing on clay courts and, in fact, he's known as the "King of Clay."

Specifying the TM

- Here's a table of *Nadal's performance* in major tournaments, from 2002-2010. As you can see, the tournament venues have different court surfaces.

Court Surface	Tournament	Career Win (%)
Clay	French Open	97.44
Grass	Wimbledon	87.87
Hard	Australian Open	83.33
Hard	U.S. Open	80.00

- His overall performance is obviously better on clay courts compared to hard courts. So, if we assessed his performance only on hard courts, our claim for his overall tennis ability would be underestimated. On the other hand, if we assessed his play on only clay courts, we may overestimate his skill.
- The point is that when assessing, we need to make sure that we set up circumstances and tasks that are sufficiently varied so that multiple sources of evidence are collected and woven into more accurate inferences.

TOC

Wrapping it Up

ECD is a powerful framework for designing and developing assessments. However, it is *not* a prescriptive model. And while it has many benefits (described on the next page) it can only lead to excellent assessments if it is thoughtfully applied.

In summary, ECD:

- Requires clear articulation of *claims* to be made about peoples' competencies
- Establishes valid *evidence* of the claim (i.e., student performance data demonstrating varying levels of mastery)
- Specifies the nature and form of *tasks* or situations that will elicit that evidence

TOC

Benefits of ECD

Flexibility. ECD provides a flexible framework to design valid and reliable assessments for various *purposes* (e.g., formative, summative), at different *levels* or grain sizes (e.g., single score or diagnostic sub-scores), and for assessing various *types* of learner attributes (e.g., conceptual understanding, dispositions, skills).

Convergency. ECD lets you aggregate all kinds of data (e.g., qualitative and quantitative) as frequently (even continuously) as you wish. This can (a) increase the reliability and validity of the assessment, and (b) move us toward fusing learning and assessment when designing for diagnostic purposes.

Transparency. Because ECD is based on evidentiary arguments, you can clearly link specific *performance data* (which are observable) to *theoretical constructs* (unobservable). Such transparency is important for accountability purposes – for all stakeholders (e.g., teachers, students, parents, administrators, policy makers).

Reusability. ECD provides a blueprint for creating assessments that can be re-used (i.e., reduce the time of preparing assessment tasks or environments). For instance, if you develop a good CM and EM for systems thinking skill, they may be used (and re-used) in various settings (e.g., simulation, game, classroom discussion, etc.).

TOC

Barriers to ECD
(or… research opportunities)

Cost. ECD takes a lot of up-front effort to get all the models right. This process consumes time and resources (e.g., consulting with experts). So one of the main barriers to scaling-up ECD is development cost. There are, however, research efforts underway to automate the acquisition of information needed to construct the competency and evidence models.

Scope. The competency model in ECD needs to be developed at just the right level of granularity to be optimally effective—for the assessment and to support learning. Too large a grain size means less specific evidence is available to determine competency, while too fine a grain size means a high level of complexity and increased resources to be devoted to the assessment.

Rubrics. Making good rubrics is hard! Moreover, even when teachers are provided with good rubrics, scoring qualitative products (like essays and online discussions) can still be subjective. So a detailed and robust coding/scoring scheme is needed that takes into account the context of the tasks and semantic nuances in students' submissions.

Task Model. When embedding assessment within dynamic learning environments, figuring out how tasks should be structured (or not) is important. Specific sequences of actions can facilitate reliable data collection, but may limit the learners' exploration of the environment. We need to find the ideal balance between exploration and structured data collection.

TOC

Helpful References on ECD

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 23*(2), 13-23.

Mislevy, R. J. (1994). Evidence and inference in educational assessment, *Psychometrika*, *12*, 341–369.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25(*4), 6-20.

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Report 632). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/r632.pdf

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the role of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.) *Item generation for test development* (pp. 97-128). Mahwah, New Jersey: Lawrence Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-62.

National Research Council (1996). National science education standards. Washington, DC: National Academy Press.

TOC

Evidence-centered design (ECD) lets you create valid and reliable assessments for important knowledge and skills!

Thank you!

Questions?

Valerie Shute

vshute@fsu.edu

TOC