The Essence of ECD

Valerie J. Shute (Florida State University)

Abstract

Before Robert Mislevy's groundbreaking work on evidence-centered design (ECD), educational measurement was too often a black box—a collection of tests that yielded scores based on opaque assumptions. ECD did more than improve test design; it reshaped the logic of assessment itself and offered a principled way to connect what we observe to what we claim about learners. This paper aims to demystify ECD for use by a broad audience. It is written in honor of Mislevy, whose work showed that if we want to measure the human mind, we must first learn how to weave.

A Metaphor: Weaving the Tapestry of Evidence

Weaving interlaces threads of different colors, textures, and strengths to form a coherent fabric. The quality of the tapestry depends on deliberate planning: which threads matter, how they interact, and how much weight each carries.

ECD works in much the same way. The "threads" are pieces of evidence—observable behaviors, products, or actions—that collectively support claims about unobservable constructs such as knowledge, skills, beliefs, values, or dispositions. These constructs cannot be seen directly; they must be inferred from patterns of evidence.

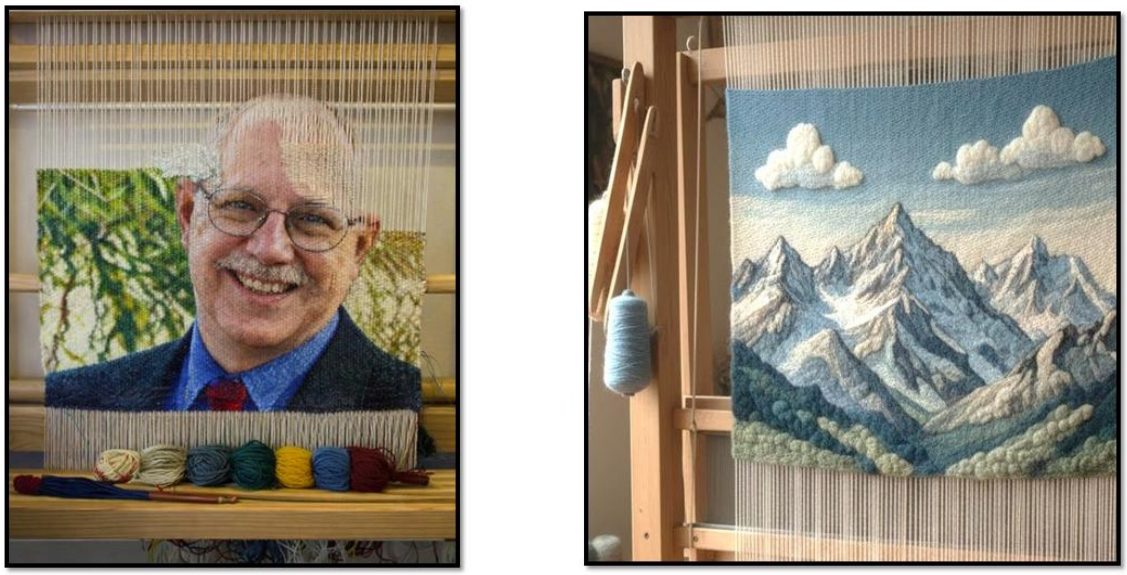

Figure 1. Examples of two tapestries-in-progress (AI generated, 2026)

Why Inference Is Central to Assessment

You cannot directly observe attributes such as creativity, critical thinking, kindness, or chess knowledge. What you *can* observe are construct-relevant behaviors—written or spoken arguments, problem-solving strategies, decisions under constraints, or actions in a game or simulation—and then make inferences about those attributes. Assessment is therefore inherently inferential. Every assessment rests on an argument: If a person can do *X* under conditions *Y*, it is reasonable to conclude that they possess attribute *Z,* to a particular degree.

ECD makes this argument explicit. Rather than hiding assumptions behind test scores, ECD requires designers to clearly specify the claims they wish to make, the evidence needed to support those claims, and the situations most likely to elicit that evidence.

Why We Really Need ECD

The demands of modern life have changed dramatically over the past century. Routine tasks are automated, information is ubiquitous, and success increasingly depends on higher-order skills such as problem solving, collaboration, adaptability, and learning how to learn—especially as AI becomes deeply embedded in everyday life.

Assessment practices, however, have been slow to change. Many systems still rely on short, decontextualized tests of recall and procedural fluency. These approaches are efficient but often fail to capture what matters most. As Robert Reich noted, standardized tests can be "monstrously unfair," branding many students as failures when they might thrive if progress were measured differently (Reich, 2000).

Generally, traditional tests are limited in several key ways:

1. *Ease over importance:* They measure what is easy to score rather than what is most valuable.
2. *Fragmentation:* They emphasize isolated bits of knowledge instead of integrated understanding or skill.
3. *Low diagnostic value:* They often provide post-hoc scores rather than actionable feedback.

ECD does not reject traditional tests. Rather, it situates them within a broader framework that can support everything from brief quizzes to complex simulations, all grounded in principled evidentiary reasoning.

Brief History, Claims, and Basic Structure of ECD

ECD originated at Educational Testing Service in the late 1990s through the work of Robert Mislevy, along with Russell Almond and Linda Steinberg (e.g., Mislevy, Steinberg, & Almond, 2003). It was built on Samuel Messick's assertion that the construct should guide task design and scoring, "*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*." (Messick, 1994, p. 20).

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

- Supports valid assessments for multiple purposes (formative and summative).
- Enables estimation of complex, dynamic competencies.
- Integrates qualitative and quantitative data from multiple sources.

- Provides transparency and accountability through explicit reasoning.

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). But ECD is not a checklist; it is a system of interrelated models. Mislevy envisioned these models as a way to describe how competencies give rise to evidence, and how that evidence, in turn, supports our claims. So, 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. In its most basic form, ECD can be described by three main models: Competency Model, Evidence Model, and Task (or Action) Model. Below is a picture showing the flow among the models then I will elaborate on each of the models in turn.

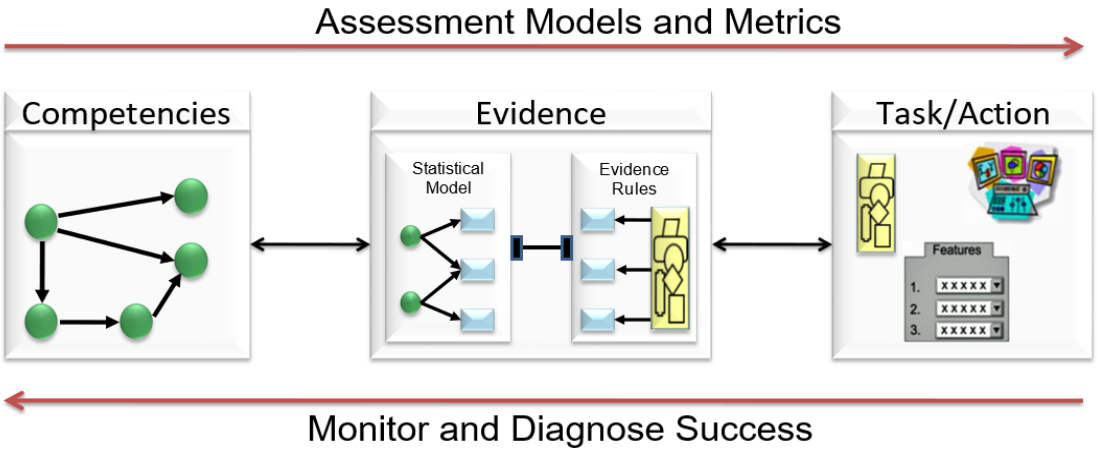

Figure 2. The three key models of ECD (adapted from Mislevy et al., 2003)

The red arrow heading left-to-right shows reasoning about assessment design (from competencies to evidence to tasks). Note that this flow is not a strictly linear process (CM to EM to TM) but rather, an iterative process where all models are considered concurrently. The arrow going from right-to-left demonstrates diagnosis and interpretation of performance. After reviewing each of the three main models in turn, I will attempt to make these come alive with a real-world example.

Competency Model (CM)

The CM answers the question: *What knowledge, skills, and attributes do we want to assess?* It consists of variables representing learner attributes at varying levels of grain size, from broad claims (e.g., overall math proficiency) to fine-grained diagnoses (e.g., difficulty solving linear equations). The term "student model" is an instantiated version of the CM, representing current beliefs about a learner's proficiency on each variable.

Evidence Model (EM)

The EM, shown in the middle section of the figure above, addresses *which particular behaviors should reveal different levels of the targeted competencies*. It has two components:

1. *Evidence Rules* (i.e., rubrics or scoring model) transform work products (e.g., answers, artifacts, action sequences) into observable scores or indicators.

2. The *Statistical Model* links observable indicators to CM variables, often using probabilistic reasoning (e.g., Bayesian networks) to update beliefs about competencies. In short, the statistical model allows evidence from multiple tasks to be combined, yielding increasingly reliable and valid inferences over time.

Task Model (TM)

Finally, the task model specifies the situations designed to elicit evidence by asking, *What particular features must a task have to elicit specific, observable evidence of a student's knowledge, skills, and abilities, and how can we systematically vary those features*? Tasks are characterized by factors such as presentation format, expected work products, and contextual variables (e.g., difficulty, time pressure).

In games or simulations, the TM is often called an action model, emphasizing sequences of learner actions and their indicators of success. Action models are especially useful for capturing rich, process-level data in dynamic environments.

Applying ECD—A Tennis Example

To illustrate ECD, consider assessing tennis skill. You begin 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. I'll show examples of that in a minute. So, what skills does a good tennis player need to have? What do novices do?
Relevant CM variables might include forehand stroke, backhand stroke, serve, and footwork. These variables only gain meaning when structured to represent their relationships.

Structuring the CM – Conceptual Depiction

Here's a possible structure of the variables (Figure 3). 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?

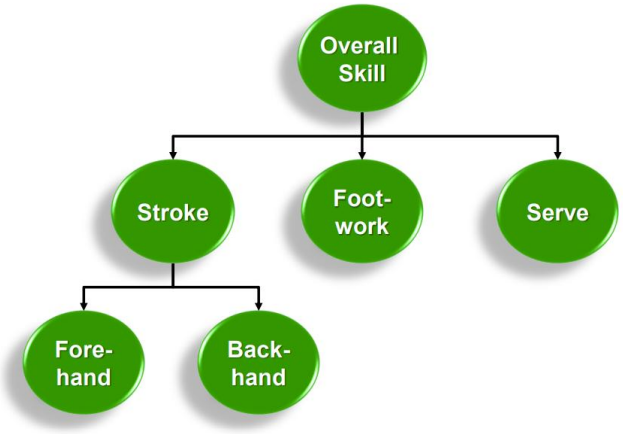

Figure 3. Key competency-model variables that relate to tennis skill

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 competency model, compared to stroke and serve.

Let's think about another situation. An aspiring tennis player named Oscar 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?

Structuring the CM – Computational Depiction

The following shows a probability distribution illustrating the expected relationships among the competency variables (see Figure 4). This represents our initial beliefs or knowledge about an unknown quantity before seeing any new data, presented in what's called a Bayesian network (or Bayes net).

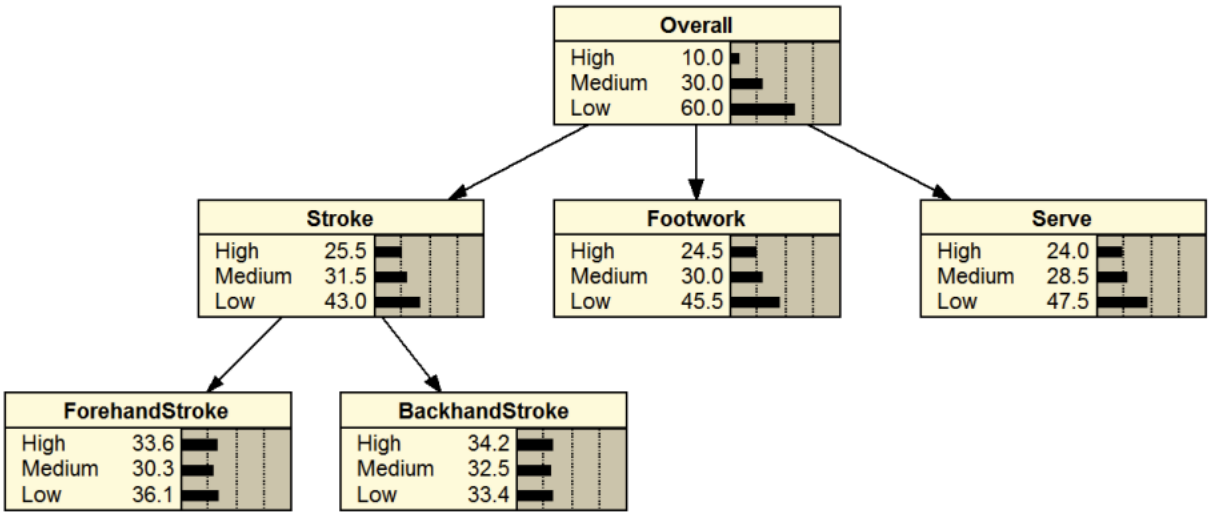

Figure 4. Prior probability distribution of our tennis-related variables

Next, Figure 5 shows two boxes greyed out, which means they've been instantiated with two observations (i.e., where we saw forehand stroke set as "high" and backhand as "low"). In this case, the Bayesian network accumulates evidence based on observations of a person's forehand and backhand strokes) and passes the information into the student model (i.e., the competency model that is specific to a student).

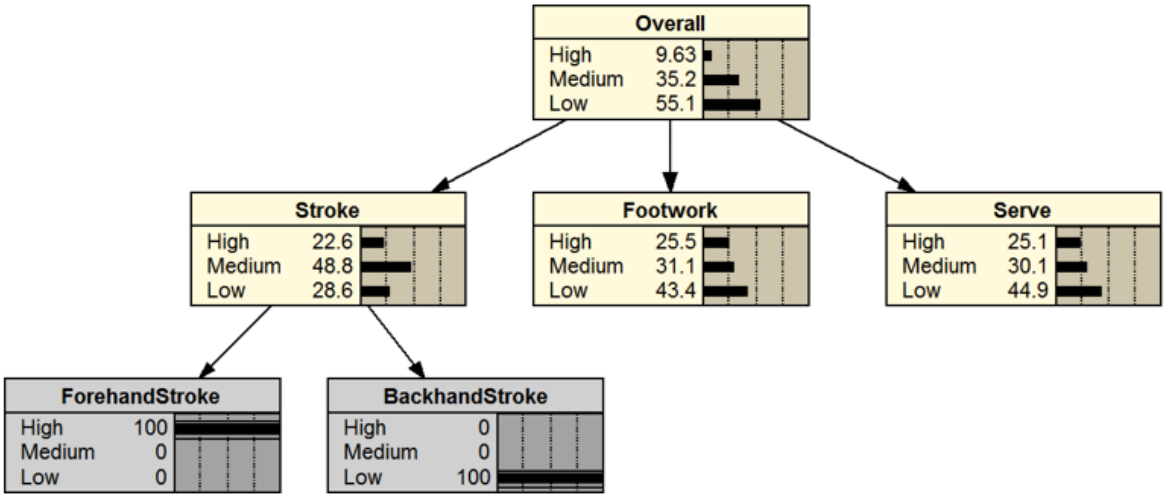

Figure 5. Posterior probability distribution of tennis variables with stroke evidence included

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.

Earlier we wondered about how we could show different degrees of influence of variables on each other. To illustrate, suppose that Oscar mastered his backhand and now demonstrated a high level of skill for both the stroke and serve variables, but a poor level of footwork skill. Figure 6 shows his current, overall tennis skill level.

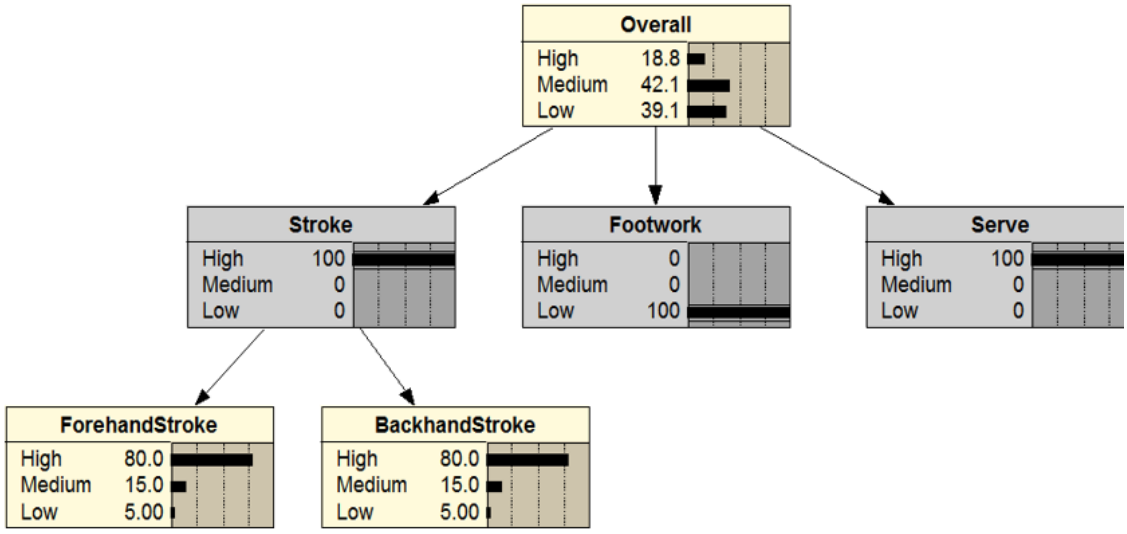

Figure 6. Posterior probability distribution with stroke, footwork, and serve evidence included

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 tennis-skill variable, which is reflected in the probabilities.

Building the EM

Now it's time to move our attention to building the Evidence Model (EM) which determines how the observed actions can be used as evidence to update the current states of the competency model variables. We need to create two parts of the EM: evidence rules and the statistical model. Keeping the same tennis example will show how evidence rules and statistical models work together.

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.

Table 1. Rubrics for tennis indicators of 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.

Having specified the scoring rules for our tennis example, we can now observe a person (e.g., Maisyn) playing tennis, and then score her forehand stroke. The highest possible score one can earn for forehand stroke is 6 (across the three indicators). Suppose Maisyn 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, like a loom in my weaving metaphor, statistically links observational data (behavioral threads) 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 (Maisyn's score on forehand stroke) / 6 (total score) = 0.50. Second, we can set cut-scores, as shown in Table 2, below:

Table 2. Cut-scores for various states/levels

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

According to the table, her score will be updated into the model as *Medium* for her forehand stroke. Figure 7 illustrates this updating process.

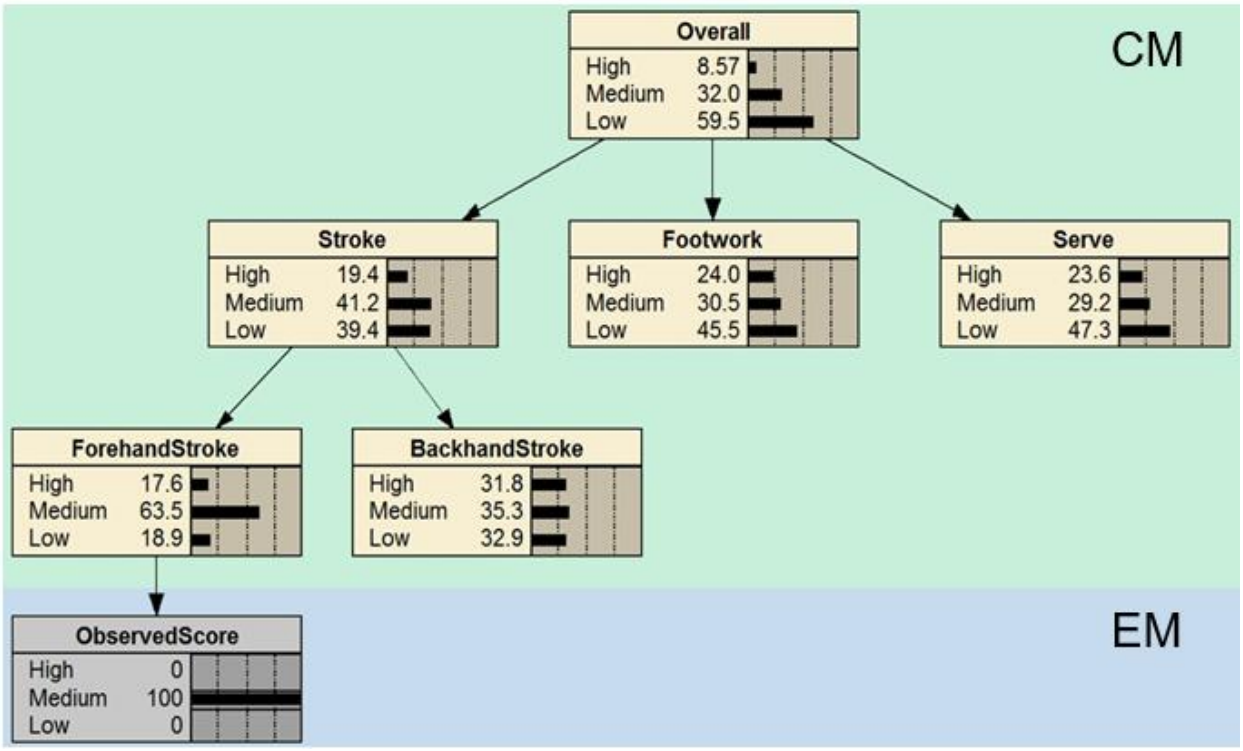

Figure 7. Bayes net update with an observed score of "medium" for forehand stroke

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 Maisyn's stroke skill is at the medium-to-low level, at this point in time and with just one observation. Additional observations (e.g., more forehand as well as backhand strokes, footwork, and serve) will further update the model and strengthen the validity of our inferences.

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.

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. 1,000-word essay on tennis
C. Let the person demonstrate his/her skills on a tennis court.

Clearly the answer is "C" as 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. 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?

Figure 8. Three different types of tennis courts: Clay, hard, and grass

Rafael Nadal grew up playing on clay courts and, in fact, he's known as the "King of Clay" (for more see: https://en.wikipedia.org/wiki/List_of_Grand_Slam_men%27s_singles_champions).

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

Table 3. Career wins for Rafael Nadal per court surface

Court Surface	Tournament	Career Win (%)
Clay	French Open	96.6
Hard	Australian & U.S. Open	83.7
Grass	Wimbledon	78.3

His overall performance is obviously better on clay courts compared to hard courts or grass. So, if we assessed his performance only on grass 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.

Wrapping it Up

In summary, ECD requires (a) clear articulation of the claims to be made about people's competencies, (b) valid evidence to support those claims (e.g., performance data showing varying levels of mastery), and (c) well-specified tasks or situations that will elicit that evidence. ECD is a powerful framework for assessment design, but it is not prescriptive. Its effectiveness depends entirely on how thoughtfully it is applied.

One of ECD's major strengths is its *flexibility*. It supports the design of valid and reliable assessments for multiple purposes (formative or summative), at different levels (e.g., single scores or diagnostic sub-scores), and across a wide range of learner attributes, including knowledge, skills, and dispositions. Closely related is *convergency*: ECD allows the aggregation of diverse data sources—qualitative and quantitative, sparse or continuous. This can improve both reliability and validity, while also enabling the fusion of learning and assessment, particularly for diagnostic purposes.

ECD also promotes *transparency*. Because it is grounded in explicit evidentiary arguments, observable performance data can be clearly linked to unobservable theoretical constructs. This transparency is essential for accountability and communication among stakeholders, including teachers, students, parents, administrators, and policymakers. Finally, ECD supports *reusability*. Well-developed competency and evidence models can serve as blueprints for multiple assessment contexts, reducing development time and allowing assessments to be adapted across settings such as simulations, games, and classroom discussions.

Despite these benefits, ECD presents several challenges—many of which also represent important research opportunities. *Cost* is a primary barrier: developing high-quality models requires substantial upfront time and expertise. Efforts are underway, however, to automate aspects of model construction. *Scope* is another challenge. Competency models must be developed at an optimal level

of granularity—too coarse, and they yield vague evidence; too fine, and they become complex and resource-intensive. *Rubrics* pose additional difficulties: even strong rubrics cannot fully eliminate subjectivity when scoring complex, qualitative products, underscoring the need for robust and context-sensitive scoring schemes. Finally, the *task model* becomes especially challenging in dynamic learning environments, where designers must balance structured data collection with learners' freedom to explore.

Using tennis as a running example helps illustrate ECD's logic. Tennis expertise is more than power hitting; it also depends on *context*, including surface type. ECD makes such contextual biases explicit and manageable. This logic also directly underpins my own work on stealth assessment (Shute, 2023). Imagine a middle-school student playing an interactive game, drawing levers and pulleys to solve puzzles. To the student, it feels like play; to the researcher, it generates rich data about evolving physics understanding (as well as other competencies like persistence, problem solving, and creativity). For instance, as students interact with the game environment, they generate a continuous stream of data captured in the game's log files. The stealth assessment filters through and analyses that data—in real-time—to identify and extract evidence related to the CM. This is the evidence identification (EI) process. The EI's output is the input data (e.g. scores, tallies) for the evidence accumulation (EA) process, which statistically updates the claims about relevant competencies in the CM (e.g. the probability of a student being low, medium or high on a given competency, like understanding Newton's second law). The more evidence a student generates during gameplay, the more accurate the estimates of competency levels. This seamless integration of assessment and learning is possible because ECD provides the scaffolding that links actions to claims. Without Mislevy's framework, stealth assessment would not exist.

In closing, Robert Mislevy was more than a psychometrician; he was a philosopher of the human story. He understood that every time we "test" a student, we are telling a story about who they are and what they are capable of doing. He wanted that story to be as accurate, fair, and helpful as possible. ECD reminds us that assessment is not about scoring tasks—it is about making justified claims about people. When we weave evidence carefully, the resulting tapestry is strong, meaningful, and worth trusting. Because of Mislevy, and through the lens of ECD, we can truly see individuals' progress, and provide precise assistance and/or applause as warranted.

Relevant References[1]

Almond, R. G., Shute, V. J., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In R. Lissitz and H. Jiao (Ed.), *Innovative psychometric modeling and methods* (pp. 163–192). Charlotte, NC: Information Age Publishing.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2) pp. 13-23, 10.2307/1176219

[1] I wrote an earlier version of this paper in 2010 with my then-students (Y. J. Kim & R. Razzouk) entitled, "ECD for Dummies" but we never published that paper. Bob did, however, read the paper and gave it two thumbs up!

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., et al. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*(2–3), 121–140.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

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.

Rahimi, S., Almond, R. G., Ramírez-Salgado, A., Wusylko, C., Weisberg, L., Song, Y., ... & Wright, E. (2024). Competency model development: The backbone of successful stealth assessments. *Journal of Computer Assisted Learning, 40*(6), 2772–2789.

Reich, R. (2000). One education does not fit all. New York Times, Op-Ed, July 11, 2000.

Shute, V. J. (2023). The history of stealth assessment and a peek into its future. In M. P. McCreery, & S. K. Krach (Eds.), Games as stealth assessments (pp. 1-23). Hershey, PA: IGI Global. Retrieved from https://doi.org/10.4018/979-8-3693-0568-3.ch001