Serious assessments in serious games

Robert Hubal
RTI International
13 April 2011

Intro to presentation

- ☑ **Interested in providing strategies that might improve how student performance is assessed in serious games.**
- ☒ **Not interested in improving the games *per se*:**
  - Take perspective of 'simulation training'.
- **Many colleagues to thank for these ideas:**
  - Brown, Byerly, Cain, Cianciolo, Clyman, Deterding, Feltovich, Frank, Hayes-Roth, Heneghan, Kenny, Parsons, Pina, Sawyer, Spira, Stanley, Strickland, Willard, Wray, Young,...

Simulation training

- **Goals:**
  - Acquire, practice, be validated on skills.
  - Transfer skills to real-world situations.
  - Cover range of situations over which to apply skills:
    - 'Coverage' is a key concept.
- **Methods:**
  - Range of immersion:
    - Everything from animated stick figures to true VR.
    - (But focus on desktop.)
  - Develop lessons or vignettes rather than games.

Serious games

- **A simulation trainer's [self-absorbed] view:**
  - The point is not to have fun. The point is to engage the user in the lesson or vignette.
  - Our users are not players. They are students.
  - We render via a game engine, and use similar tools as game developers:
    - But training is not a game, it has direct purpose.
    - It may be enjoyable, and may need a storyline and gamelike play, but those features are in essence secondary.

Serious games

- **A[n open minded] simulation trainer's view:**
  - Simulation trainers have a lot to learn from game designers:
    - Narrative.
    - Theme, thematics (e.g., background sounds).
    - Better measures of engagement.
  - It only makes sense to take advantage of existing content, capacity, and experience.
- **But... the intent here is to see if game designers have anything to learn from simulation training experience:**
  - And specifically to focus on *assessment*.
  - And even more specifically, *performance* assessment.

Performance assessment

- **Demand is to assess students' capabilities or learning of skills.**
- **Do so in a 'situated' environment.**
- **Must be actionable:**
  - What does the student know how to do?
  - What can the student do in what context?
  - Where does the student go next?

(Frank et al., 2005)

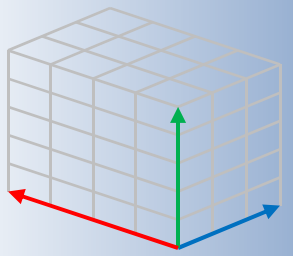
Situated assessment

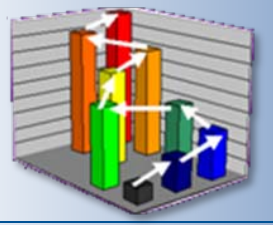
- **Learn / assess by doing:**
  - Makes sense in most situations where procedures and strategies hold.
- **Address 'imperfect conceptual models':**
  - Present faults or adverse lessons/vignettes needing to be addressed that are not obvious.
  - Dynamic performance measures of critical tasks.
- **Move away from non-interactive (surveys) and non-distributive (hands-on).**

(Hubal, submitted)

Training course

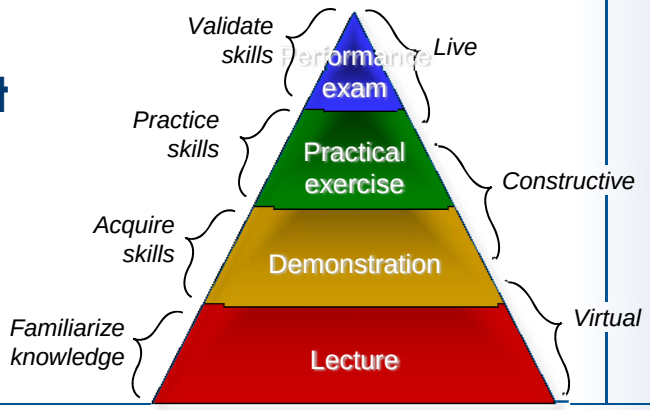
- **Organized into lessons or vignettes:**
  - Learning progression.
  - ⋆ Lessons/vignettes address specific performance criteria.
  - Assessment determines student GO/NOGO.
- **Lessons or vignettes can be skipped if student already knows material as determined by some type of initial assessment.**
- **Partial ordering in lesson/vignette sequence, but students not forced to comply with ordering:**
  - Recommended or remedial sequencing is based on analyses of student performance.
  - ⋆ Sum total of lessons and vignettes must cover learning space.





Learning progression

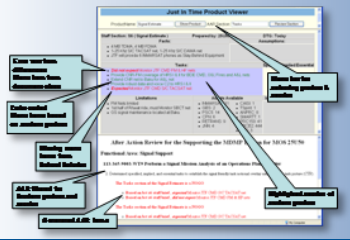
1. **Gain knowledge about components or events or procedures; assess through tests.**
2. **Acquire and practice skills:**
   - Start by learning "school solution" or best-practice procedures, often lock-step.
   - Gradually move to freer play.
   - Learn by doing, multiple lessons/vignettes with different 'fault' conditions, reach-back to supporting materials.
   - Tutoring system reacts to differences between student actions and performance criteria.
3. **Test on performance of skills to established standards within set conditions.**



(Hubal et al., 2000)

Tutoring

- **Remediation & forward recommendation.**
- **Design of lessons or vignettes:**
  - Set of lessons/vignettes needs to encompass competencies.
  - Thus, need to have theory defining competencies.
  - Need to be realistic (engaging).
- **"Representative" tasks:**
  - Need to be realistic (relevant).
  - Consider context.
  - The experience in location A should equate to the experience in location B.
  - Describe research as about competent performance, not expertise.
- **How to define levels of difficulty such that advancement through them reflects increased skill development.**



Outline for rest of talk

- **Three broad groups to consider during design/development:**
  - Characteristics of the task.
  - Characteristics of the student.
  - Characteristics of the domain.

(Hubal, 2005)
(Hubal, 2008)

Task characteristics

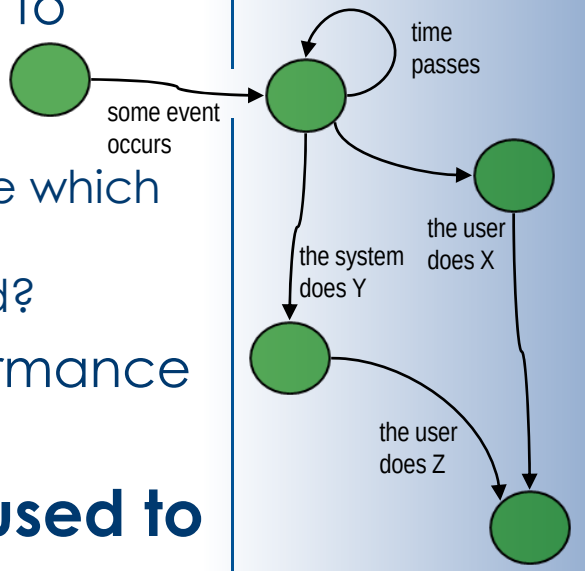
- **Those design decisions that define game entities that can be manipulated to help the student be adaptive to real-world environments.**
- **Six categories:**
  - Temporal factors,
  - Sequencing aberrations,
  - Effect of incomplete information,
  - Variability of tools and their functions,
  - Variation in the actors in the environment,
  - Environmental noise and distraction.

Temporal factors

- **Time fidelity:**
  - Some task timing should be mimicked:
    - Triage, de-escalation (through negotiation), and IED defusal are examples.
    - Hydraulics startup and watching paint dry are not:
      (Focus on process and speed it up.)
    - But slow it down early in learning.
- **Time pressure:**
  - Depending on skill progression, ignore or require temporally accurate response:
    - Consider the student who is learning to negotiate.
  - Sometimes *imposed* time pressure is desirable:
    - The psychologist's speed/accuracy tradeoff can inform learning progress.

Sequencing aberrations

- **Errors of omission/commission/sequence:**
  - Complex game states may be needed to track process (i.e., user actions):
    - When must an action occur?
    - After which action or condition and before which other? In parallel with another?
    - May it be skipped, or repeated, or inserted?
  - ★ Consequently complex dynamic performance measures are managed.
- **In game development, "gating" is used to control flow:**
  - Or else nonlinearity in emergent gameplay is accommodated.



(Frank & Hubal, 2007)
(Guinn & Hubal, 2006)

Incomplete information

- **Students may make decisions based on partial information. Examples are:**
  - Poor differential diagnosis of medical condition.
  - Failure to disconfirm due to bias.
  - Lack of full awareness of function of equipment.
- **During acquisition, intervene at point of taking wrong branch:**
  - (See network on prev. page.)
- **During practice, intervene at teachable moment that usually comes later:**
  - (When student – "uh oh" – realizes the impasse.)
  - Requires ongoing student modeling.
  - For adaptive assessments, requires maintenance of dynamic performance criteria.
  - Asking student for an explanation can make missing information more apparent.

(Hubal & Frank, 2008)
(Kizakevich et al., 2003)

Variability of tools

- **As with weapons in games, what 'tools' are in the student's arsenal influence activity:**
  - Analogy is with different first responder trauma kits.
- **Tools obviously should work only when used appropriately:**
  - The right tool (stethoscope, multimeter, wrench) applied to the right location.
- **Game developers already consider the min/max player.**
  - Change the challenge dynamically for the student.
  - But reward appropriately.

Variation in actors

- **Presence, number, and behavior of NPC's can influence performance:**
  - Across lessons or vignettes, a character can play one or different roles:
    - Roles can make the task harder (e.g., by introducing biased responses).
    - Different characters (e.g., differing in appearance) can take the same role in different instances.
  - Character actions supported by behavior models:
    - Emotions.
    - Knowledge.
    - Social graces.
    - Animations.
    - Physiology.

(Hubal et al., 2003)
(Deterding et al., 2005)

Environmental distraction

- **Some students, in some situations, are affected by 'noise':**
  - Early in the learning progression, minimize background activity:
    - (Not just the task itself.)
  - Bring in background activity to:
    - Enhance realism.
    - Entice biased actions.
    - Increase task difficulty.

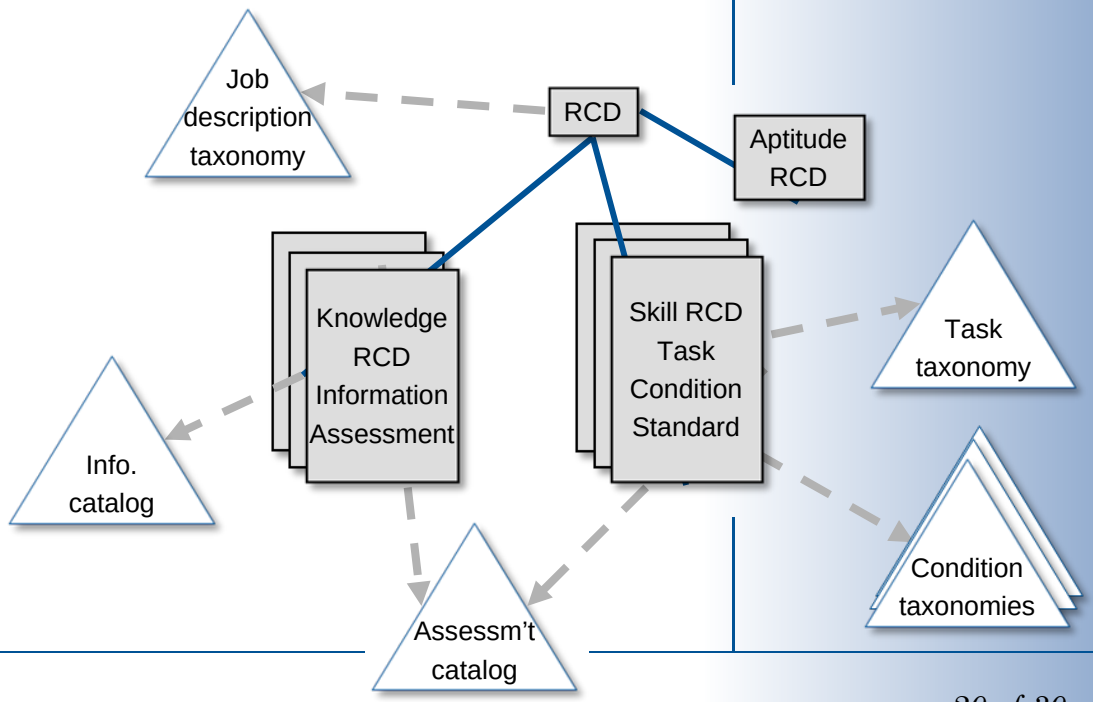
(Mills & Hubal, 2001)
(Hubal & FitzGerald, submitted)

Student characteristics

- **Those design decisions that define game entities that can be manipulated to help the environment be adaptive to real-world students.**
- **Four categories:**
  - Reusable competency definitions,
  - Motivation to learn,
  - Performance levels,
  - Demographics & traits.

Competency

- **Measure using established tasks, conditions, and standards.**
- **Generalize whenever possible:**
  - Use known constructs.
  - Use representative and comprehensive lessons or vignettes.



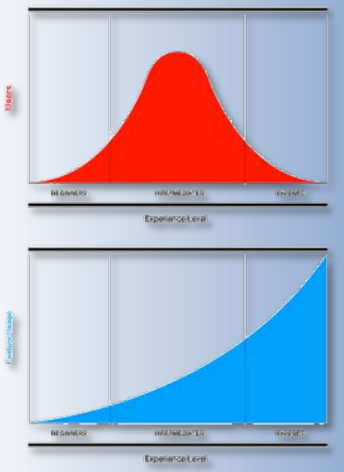
(Frank et al., 2007)

Motivation to learn

- **Motivation in games is largely fun through challenge:**
  - Games present opportunities for harder challenges as the player's skill level increases.
- **Motivation in simulation training is to learn:**
  - Can be internal or external.
  - Challenge through learning progression mirrors that for games.
- **Form of learning influences design and thus assessment:**
  - Implicit or inductive learning makes the gameplay prominent, but performance assessment complicated.
  - Explicit learning makes the narrative critical (for engagement) but performance assessment 'invasive'.

Performance levels

- **What are different levels of performance? How many (if any) different levels exist?**
  - Expertise ⇔ mastery ⇔ proficiency ⇔ familiarity.
  - Different types of content experts.
- **Rather than try to assess the player's skill level, a game might maintain an idea of how skilled it expects the player to be by a certain point.**
  - But this approach does not work when specific performance criteria are measured.



(Hubal, 1996)
(Hubal, 2009)

Demographics, traits

- **Apparently there are different types of people in this world:**
  - Nothing of value to say wrt gender, ethnicity, age, personality type:
    - E.g., have never found a consistent effect on engagement.
  - But there may be effects on performance for certain tasks under certain conditions:
    - Some individual differences research suggests there are effects on sustained attention to psychomotor tasks:

      Gamers seem to do well on these tasks.
  - Classic use by game developers of Bartle's types (achievers, explorers, socializers, killers).

(Hubal et al., 2010)
(Bartle, 1990)

Domain characteristics

- **Those design decisions that define game entities that can be manipulated to adapt to features of the context in which tasks take place.**
- **Four categories:**
  - Critical tasks & performance measures,
  - Red screen alerts,
  - Ill-structured or wicked domains,
  - Violence.

Performance measures

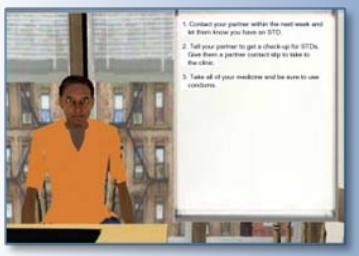
- **Ongoing and after-the-fact.**
- **Categorical – actions might be correct, incorrect, don't care:**
  - Incorrect actions placed into predefined categories.
  - What happened (performance measures).
  - Why it happened (performance measure criteria).
  - How it happened (student actions).
- **Based on student actions and simulation state, decide whether and how to intervene.**
- **Evaluate overall progress through training course as well as through individual lessons/vignettes.**

Red screen alerts

- **Game over:**
  - SNAFU in a big way.
- **During simulation training:**
  - Use sparingly, there must be a serious training message.
  - Works only within active monitoring:
    - (Not after-action reviews – they are too late.)
  - As with games, punish by loss of time.
- **Keep the message simple.**

Wicked domains

- **Those that have:**
  - Few best practices.
  - Few established metrics:
    - Or those requiring atypical measures such as nonverbal.
  - May require a large number of lessons or vignettes to 'cover' the space.
- **Examples:**
  - Establishing trust with a pediatric or schizophrenic patient.
  - Establishing a provincial reconstruction team or supporting stability ops.
  - Learning to discuss sensitive topics.



Violent domains

- **It is known that violent games can contribute to real-life violence and aggression.**
- **But content is just one component when assessing risky decision-making:**
  - Can also assess using carefully constructed lessons or vignettes.

(Greitemeyer & McLatchie, 2011)
(Hubal et al., 2008)

Implications

- **Simulation training ≠ gaming:**
  - But the tools game developers use are important to simulation training.
  - And motivation to learn ≠ motivation to play, but they are not mutually exclusive.
    - They both involve challenge, and they both can be fun.
- **Get in the student's head:**
  - How do I *demonstrate* (not describe) what I've learned?
  - How does my performance show what I've not learned to perfection?
  - Where am I taken next to learn more?
- **Create lessons or vignettes based on domain-relevant constructs, using representative (or transferable) tasks, tailored to individual students:**
  - Create assessment of performance of tasks under specified conditions to set standards.

Questions?

- **Contact:**

Robert Hubal, PhD

Senior Research Psychologist

RTI International

919/541.6045

rhubal@ rti.org