The Challenge of Game, Learning, and Assessment Integration

Abstract: The potential of games as learning and assessment tools can be met only if replicable methods for aligning game play with learning standards and formative assessment objectives can be developed. In addition, we must be able to adapt traditional measurement methods to new types of activity and data. The goal of this symposium is to present key aspects of the interplay among games, assessments, and learning based on a large-scale game development project. Using examples from SimCityEDU as illustration, design, data tools, psychometric modeling, and evaluation studies will be discussed.

Presentation 1: Designing Fun Learning and Assessment Games

Erin Hoffman and Michael John
Institute of Play

SimCityEDU

Before the discussion of design, a brief summary of the developed game will provide context for the work described throughout the paper. *SimCityEDU*, based on the popular *SimCity* commercial game, asks players to solve problems facing a city, generally requiring them to balance elements of environmental impact, infrastructure needs, and employment. The game scenarios are designed to assess systems thinking. Often named on lists of 21st century skills, systems thinking is also a cross-cutting concept in the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Essentially, it is the understanding of how various components of a system influence each other. Table 1 presents a summary of the systems thinking learning progression used in *SimCityEDU*.

Level 1 - Acausal The player is not reasoning systematically about causes and effects.
Level 2 - Univariate The player tends to focus on a single causal relationship in the system
Level 3a - Early Multivariate The player has considered multiple effects resulting from a single cause
Level 3b - Multivariate The player has considered multiple causes in relation to their multiple effects
Level 4 - Emergent Patterns The player attends to and intervenes on emergent patterns of causality that arise over time

Table 1: Systems Thinking Learning Progression from SimCityEDU

In the Jackson City scenario, the player enters the city (Figure 1a) and is told that residents seem unhappy and are leaving the city. Interaction with the Sim characters reveals that they are having trouble with air pollution. Players can explore data maps that show which buildings are polluting (Figure 1b), how power is dispersed in the city, and how various areas are zoned. Players discover that coal plants are the biggest cause of pollution in the city. However, coal plants also provide much of the power in the city. Power impacts both resident happiness and jobs (unpowered businesses close down).

Figure 1: View of Jackson City from overhead (a) and via pollution map (b).

In the game, players can bulldoze buildings, place new power structures (wind, solar, or coal generated), build new roads to expand their city, and zone and dezone residential, commercial, and industrial areas in order to achieve their goals. They can monitor the effects of their actions on pollution and jobs with the on-screen thermometers. The players' actions are captured and provide evidence in a Bayesian Network (BN) trained to provide probability estimates that individuals at each level of the systems thinking would be observed engaging in each activity, activity grouping, or sequence.

In the design and development of game-based formative assessment, the primary challenge is how to create the compelling experiences that exist in good video games with the learning and assessment goals that are valued by parents, educators, and policy makers, and kids. Creating a successful videogame without learning and assessment is hard – videogame developers claim the success rate for games that go into production is worse than the success rate for movies. Combining commercially successful learning and assessment in a technology-based environment without gaming is also hard. Combining the affordances of games with academically valued learning and assessment creates a design problem with a new combination of criteria and constraints – and there are not that many truly successful examples.

This project used a variety of approaches to solving the larger "games for education" design problem. Our focus is formative assessment, which we see as a central connection point among games, learning, and assessment. The challenge we are embracing is to grow a studio that develops successful examples of game-based formative assessment, along with methods and machinery for others to use and build upon. Our approach is to bring together experts in commercial game design, researchers and developers in learning, and assessment development to create solutions to this jointly-constrained design problem.

Figure 2 presents goals and constraints across game, learning, and assessment design that must be aligned in a game-based formative assessment for it to be successful as a game, promote learning, and provide useful evidence and inferences about what students know and can do. At the macro-level, the meaning of the game, the learning goals (knowledge and skills to be acquired), and the constructs to be assessed must support each other. At the micro-level, deeper into the design, the game mechanics (actions), learning activities, and evidence collected for assessment must connect to each other, and also be in concert with the macro-level. Taken together these levels of the three strands and their intertwining define the product. We employ an evolving co-design process we are terming evidence centered game design (ECgD) throughout design and development.

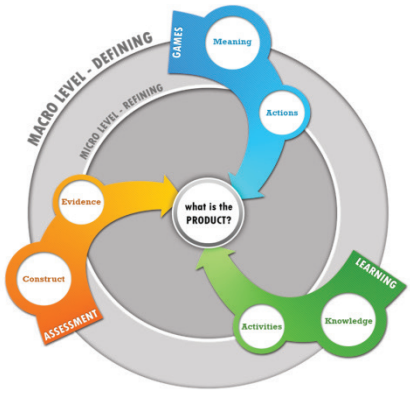

Figure 2: Integration of game, learning, and assessments goals and constraints

Meeting all of these goals in a single product has meant iterating and refining the processes by which the Lab approaches the games' designs. While this process is incomplete, the team has alighted on two key methods of game design, which are used in concert on its products.

Method 1: Emotion to Mechanics

- Key tenet of this method: begin with a desired emotion or insight, and design mechanics and aesthetics to invoke this emotion.
- Example from commercial games: *Journey.* Explores the emotion of connection (vs. loneliness) through novel online game play technique.
- Example from research team: *SimCityEDU.* Explores the fact that resolving environmental issues requires engaging difficult and unstable systems, using systemic and semi-opaque mechanics.

Method 2: Competency to Gameplay

- Key tenet of the method: begin with a deconstruction of a competency, and then find aligned mechanics which can be woven together into a gameplay experience.
- Example from commercial games: *FIFA Soccer.* The competency of Soccer can be deconstructed into *ball control, pass,* and *shoot;* the computer game in its core mode uses *joystick*, *button A*, *button B* to emulate this competency in clear mechanics.
- Example from research team: The competency of argumentation can be deconstructed into *evaluate evidence*, *form Claim-Evidence (C-E) pair,* and *critique an opponent's C-E pair;* the digital game thus has the core mechanics of *Collect Evidence*, *Construct Core,* and *Argument Attack*.

These ideas inform the discipline that GlassLab uses throughout the process of iteration called for in ECgD. The commercial games listed are each the product of years of production and millions of dollars of investment. These ideas thus are not blueprints but rather are focusing mechanisms to keep the team on a narrow path while the game is iterated around it.

Presentation 2: Uncovering the Unexpected

Kristen DiCerbo and Shonte StephensonPearson
Institute of Play

In describing the core elements of assessment delivery, Mislevy, Steinberg, & Almond (2002) suggest the importance of Evidence Identification (EI) as the process of applying scoring rules to learner work products to produce scores. However, when the work product is a log file of actions a student has taken in a game, it is less clear how to identify the scoring rules, much less apply them. We generally have only weak theories or initial hypotheses about what game behavior relates to the things we are interested in measuring.

Exploratory data analysis (EDA) is a conceptual framework with a core set of ideas and values aimed at providing insight into data, and to encourage understanding probabilistic and nonprobabilistic models in a way that guards against erroneous conclusions (Behrens, DiCerbo, Yel, & Levy, 2012).

Tukey used the analogy of data analyst as detective to describe EDA. Detective work is essentially exploratory and interactive, involving an iterative process of generating hypotheses and looking for fit between facts and the tentative theory or theories. Detective work and EDA are both essentially bottom-up processes of hypothesis formulation and data collection.

Tukey (e.g., 1986) did not consider methodology as a bifurcation between exploratory and confirmatory, but considered quantitative methods to be applied in stages of exploratory, rough confirmatory, and confirmatory data analyses. In this view, EDA is aimed at the initial goals of hypothesis generation and pattern detection following the detective analogy. Rough confirmatory data analysis is sometimes equated with null-hypothesis significance testing that is often what is taught in statistics courses. Strict confirmatory analyses involve the more sophisticated testing of specific relationships and contrasts that is actually less common in research practice.

In the creation and revision of evidence models for *SimCityEDU*, the team used a set of tools suggested by EDA that allow researchers to become intimately familiar with their data. One way to categorize and explain the tools is through four Rs: Revelation, Re-expression, Resistance, and Residuals.

Revelation refers to Tukey's (1977) statement that "The greatest value of a picture is when it forces us to notice what we never expected to see" (p. vi). Graphics are the primary tool for the exploratory data analyst. Graphical representations can display large amounts of information using relatively little space and expose relationships among pieces of information better than other representations. Here we are talking not about visualization for public display, but for finding patterns in relationships. Tools for this include things like boxplots and scatterplot matrices. In addition, interactive graphics form an important part of the toolbox, allowing the analyst to explore relationships with a few clicks.

Using a scatterplot is one of the many tools we can use to detect patterns. For example, let's consider the left of figure 3 below. It aims to explore air quality scores after a player's first and second attempt. By first glance, it may appear as though this scatterplot is just a random scattering of dots without pattern. A closer look, however, reveals that there are possibly groups of players within the scatter. As identified by the circles in the right of figure 3, we can state the hypothesis that there are three clusters emerging from the data and they appear ordered from low to high. We might also form a competing hypothesis that there are actually five groups and an outlier (the bottom picture), separating, for example, the group that starts high and stays high from the group that starts high and decreases on the second attempt. In this case, we generate new hypotheses for each additional cluster pattern we suspect, and then conduct confirmatory analyses to test these competing theories.

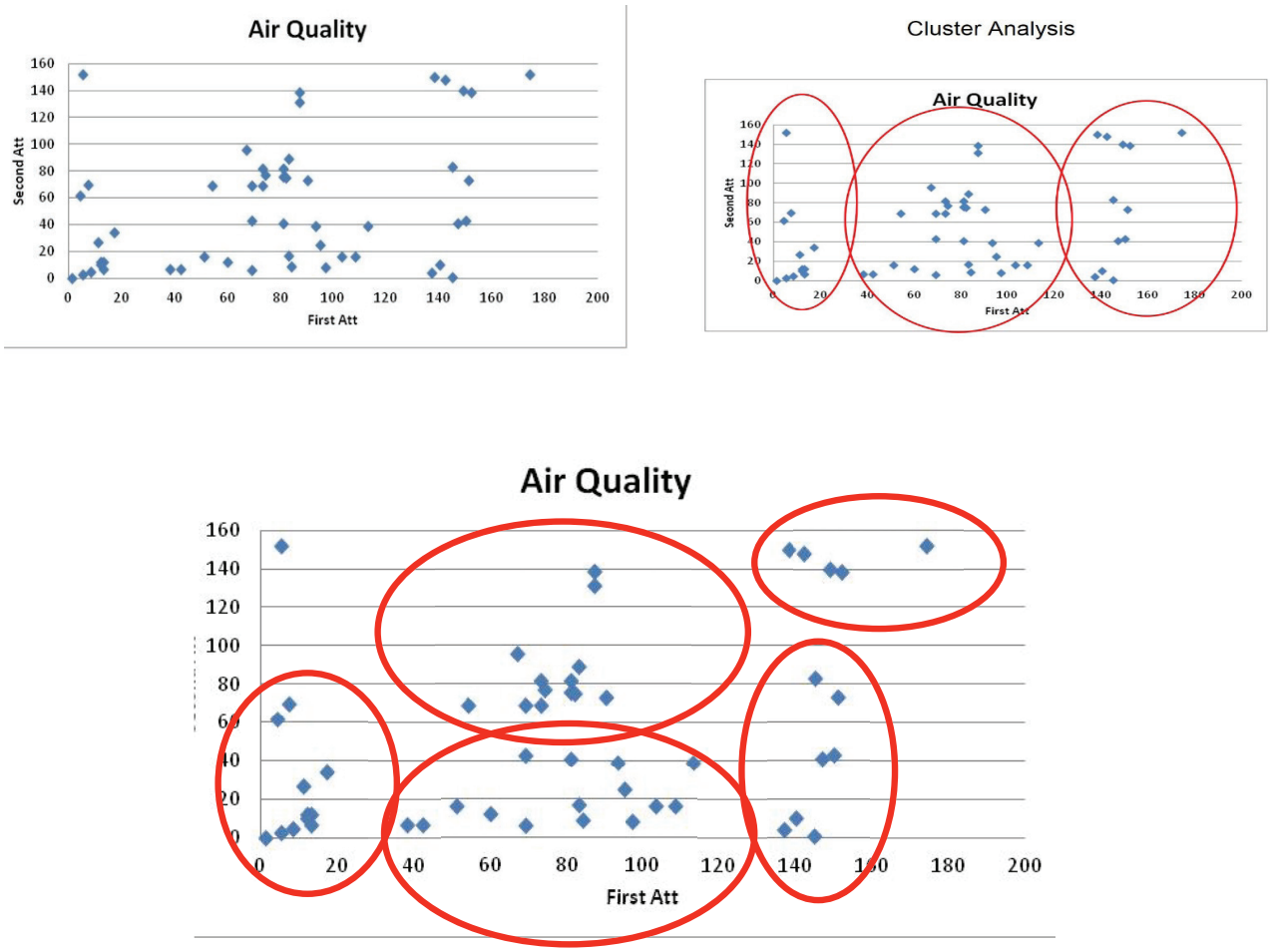

Figure 3: Use of basic scatterplots to look at patterns in data.

Data often come to the exploratory data analyst in messy, nonstandard, or simply not-useful ways. This may be overlooked if one assumes the data distributions are always well behaved, or that statistical techniques are sufficiently robust that we can ignore any deviations that might arise, and therefore skip detailed examination. In fact, quite often insufficient attention is paid to scaling issues either in advance, or during the modeling phase, and it is not until the failure of confirmatory methods that a careful examination of scaling is undertaken. Addressing appropriate scaling in advance of modeling is called re-expression and is a fundamental activity of EDA.

Resistant methods are methods that are not easily affected by extreme or unusual data. In general, there are three primary strategies for improving resistance. The first is to use rank-based measures (e.g., the median) and absolute values, rather than measures based on sums (e.g., the mean) or sums-of-squares (such as the variance). For measures of spread, the interquartile range is the most common. The second general resistance building strategy is to use a procedure that emphasizes more centrally located scores, and uses less weight for more extreme values. This category includes trimmed statistics in which values past a certain point are weighted to zero, and thereby dropped from any estimation procedures. A third approach is to reduce the scope of the data one chooses to model on the basis of knowledge about extreme scores and the processes they represent.

George Box (1976) succinctly summarized the importance of aligning model choice with the purpose of the analysis writing: “All models are wrong, some are useful” (p. 3). Residuals allow us to understand how our models are wrong. This emphasis on residuals leads to an emphasis on an iterative process of model building: A tentative model is tried based on a best guess (or cursory summary statistics), residuals are examined, the model is modified, and residuals are reexamined over and over again. In this way, throughout the analysis of the SimCityEDU log file data, hypotheses about evidence for systems thinking in game play were developed, tested, and modified.

Presentation 3: Psychometrics Meets Fun

Yue Jia, Robert J. Mislevy, Johnny Lin, Andreas Oranje
ETS

Educational assessments have been used for more than a century to gather information on what students know and can do and to guide and evaluate their progress in learning. The focus of these assessments has been on fairness, reliability, and validity of the results. In contrast, games are focused entirely on the experience of the interactions themselves and as such have the capability to engage and motivate students in immersive environments, providing goals, challenges, and rewards. Game-based assessments, such as SimCityEDU, drawing on the strengths of games, learning, and assessments, offer a great opportunity to guide learning, and to measure higher order, constructive, and interactive skills that are difficult to capture with traditional assessments (Gee, 2007, 2008).

In familiar educational assessments, a wide range of measurement models and scoring procedures have been developed to support reasoning from evidence to a given purpose or trait. However, as described in DiCerbo and Behrens (2012), most psychometric methods for educational assessments apply to what they call “the digital desert”: relatively sparse, self-contained, bits of evidence gleaned from answers to multiple-choice questions and raters’ scores for students’ essays. For psychometrics to take advantage of the richness of interactions (and associated data captured) in GBAs, we need to adapt, and where necessary extend, psychometric concepts and familiar measurement methods to new types of activity and data, such as model the multiple and interacting aspects of knowledge and skill and the dependencies among actions across time points.

As a promising psychometric framework, this paper focuses on the application of Bayesian inference networks, or Bayes Nets (BNs) for short, to GBAs. BNs are a class of models that have been developed to support probability-based reasoning as a means of transmitting complex observational evidence through a network of interrelated variables (Jensen, 1996; Murphy, 1998). BNs belong to the family of probabilistic graphical models. They can be applied as psychometric models by defining observable variables that depend on unobservable (latent) variables (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996). In a GBA, the features of situations that engage players and the actions they can take to advance play are now viewed as situations that evoke thinking through the targeted learning, and their actions provide evidence about their capabilities. It is here that game design connects with assessment design (for which we draw on the “evidence-centered” assessment design framework, or ECD; Mislevy & Haertel, 2006).

A BN consists of a set of variables (referred to as nodes) with a set of directed edges (represented by arrows) from parent nodes to child nodes indicating conditional dependence relationships between the corresponding variables. Nodes in a BN may take discrete or continuous values. In the discrete case, the directed edge represents a conditional probability table (CPT) for values of the child node, given values of the parent node. More general conditional probability distributions (CPD) are possible, such as Gaussian linear CPDs (Koller & Friedman, 2009). The graphs are acyclic in that following the directional flow of directed edges from any node it is impossible to return to the node of origin (Jensen, 2001; Pearl, 1988). By representing the variables of the model as nodes in the graph and using edges in the graph to represent patterns of dependence and independence among the variables, the model serves as a bridge between cognitive scientists, test developers and psychometric experts. In particular, “hidden” nodes corresponding to aspects of students’ knowledge, skills, and strategies, play the role of person variables in psychometric models, and aspects of performance are modeled as functions of these variables and features of game situations. The general framework can be adapted to a wide range of data types, continuous interactions, dependencies, and learning that are found in serious games.

In the presentation, we will first present the opportunities and challenges associated with using BN with GBAs.We then provide examples of building up BNs as part of an assessment engine for SimCityEDU. We will show how the cognitive models (in ECD speak: student and evidence models) and claims that are made based on those translate to fragments of the BN that was utilized.

Presentation 4: Do Assessment-based Games Keep Their Promises? : Evaluating Underlying Principles, Game Constructs and Classroom Use

Geneva D. Haertel, Terry P. Vendlinski, Britte H. Cheng, John Murray
SRI International

Technology can situate students in settings that simulate real-world environments and accumulate direct evidence of student thinking, problem solving, and understanding (Vendlinski, Chung, Binning, & Buschang, 2011). Recent advances in measurement and statistical modeling support the integration and interpretation of this accumulated evidence to yield valid inferences about student performances (Mislevy & Haertel, 2006). In theory, the Evidence Centered Design (ECD) framework accommodates these recent advances in the learning sciences (both cognitive and situative), technology and measurement. Given that ECD will frame both game, and assessment design and delivery in this project, we explored what constructs assessment-based games developed using ECD are measuring and how these game-based assessments and their associated data are actually being used in classrooms.

Since the infusion of ECD into the game and assessment design process may necessitate changes to both proven game design and student assessment methods, we documented the use of ECD in the design process in detail. Four principles of ECD, which apply to the design of game-based assessments, will be presented. These principles were derived from ECD theory and practice as well as from the experiences and reflections of the team of game and assessment designers as they designed, developed, and modified iterations of the first game (SimCityEDU). Next we will present the results from two empirical studies of SimCityEDU. The first of these studies, cognitive labs conducted on 55 middle school students, provided evidence of the constructs a student used to play the game. In particular, we employed this method to collect evidence that a student used systems thinking in the SimCityEDU game as this was the construct intended to produce successful game-play.

The next study was conducted to ascertain the classroom conditions and contexts, and instructional practices around which the SimCityEDU game-based assessment could be successfully implemented. The data collected for this study included: structured teacher interviews (after professional development and after nine days of classroom use); multiple classroom observations of each teacher; and student surveys of attitude and usage. This study of 10 middle-school science classrooms documents the range of teacher practices and student activities and attitudes associated with successful implementation.

In this presentation, we will first discuss the extent to which ECD principles were implemented in the game and assessment design process: 1) how was the domain of interest (systems thinking) organized and bounded; 2) how were the student, evidence and task models defined; 3) how was the evidentiary argument (student, evidence and task models) represented and iterated over time; 4) what design principles were implemented in the game design process (e.g., avatars, progress menus, game pacing, thermometers and other feedback mechanisms, data and other game views, narrative, looping, rewards); 5) results of cognitive lab studies about whether the construct of systems thinking was elicited during the SimCityEDU game; and 6) results of the study to document the classroom conditions, instructional practices, in particular, formative assessments and learning contexts that surround successful implementation of the SimCityEDU game. Finally, the initial results of on-going analysis of game-event logs will be presented.

References

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. Applied Psychological Measurement, 23, 223-237.

Box, G. E. (1976). Science and statistics. Journal of the American Statistical Association, 71, 791–799.

DiCerbo, K. E. & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.) Computers and their impact on state assessment: Recent history and predictions for the future (pp. 273-306). Charlotte, North Carolina: Information Age Publishing.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. Computers in Entertainment (CIE), 1(1), 20–20.

Jensen, F. V. (1996). An introduction to Bayesian Networks. London, UK: UCL Press.

Jensen, F. V. (2001). Bayesian networks and decision graphs. New York: Springer-Verlag.

Kim, Y. J., Wardrip, P., Stokes, B., Ingram-Goble, A., Shapiro, R. B., Almond, R., & Gee, J. (2012). ECDemocratized: The democratization of educational assessment. Proceedings of Games+Learning+Society 2012, 355-362. Retrieved from: http://press.etc.cmu.edu/files/GLS8.0-proceedings-2012-web.pdf

Klopfer, E., Osterweil, S., and Salen, K. (2009). Moving Learning Games Forward: Obstacles, Opportunities & Openness. Retrieved from http://education.mit.edu/papers/MovingLearningGamesForward_EdArcade.pdf.

Koller, D. & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. User-Modeling and User-Adapted Interaction, 5, 253-282.

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

Murphy, K. P. (1998). Inference and learning in hybrid Bayesian networks. Berkeley, CA: University of California

Pearl, J. (1988). Probabilistic reasoning in intelligent systems. San Mateo, CA: Morgan Kaufmann.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Tukey, J. W. (1986). Data analysis, computation and mathematics. In L. V. Jones (Ed.), The collected works of John W. Tukey: Vol. IV. Philosophy and principles of data analysis: 1965–1986 (pp. 753–775). Pacific Grove, CA: Wadsworth. (Original work published 1972)

Vendlinski, T. P., Chung, G. K. W. K., Binning, K. R. & Buschang, R. E. (2011). Teaching rational number addition using video games: The effects of instructional variation (CRESST Report 808). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).