An Introduction to Creating Video Games for Measurement: From Design to Analysis | $90

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Elizabeth Redman, UCLA CRESST; Gregory Chung, UCLA CRESST; Tianying Feng, UCLA CRESST; Jeremy Roberts, PBS KIDS

Participants will learn about considerations integral to the creation of videogames for measuring player learning, including the affordances of different game mechanics and design choices on gameplay data and how to derive meaningful indicators from gameplay data. We will use a variety of games to demonstrate how particular game mechanics impact the collection of gameplay data, the analyses that consequently can be performed with those data, and what they can reveal about player learning. This introductory session is for people interested in learning more about designing or using games for measurement purposes. It will not cover advanced statistical modeling or data mining.

The training session will have three parts. Part 1: Identifying Game Mechanics for Measurement will offer an overview of the relationship between game design and gameplay data. Part II: Extracting Meaningful Events and Indicators from Gameplay Data will offer hands-on experience with the critical analytical process involved identifying important events and deriving indicators. Part III: Examples of Indicators and Analyses of Gameplay Data will focus on basic data analysis approaches that can be used to make sense of gameplay data. Participants should bring a laptop, tablet, or smartphone to access games for the hands-on activities.

An Introduction to the Theory of Standard Setting | $90

Thursday, April 11th | 8:45 AM – 12:45 PM EDT

Mark Reckase, Psychometric Solutions

This training session is designed to give attendees a thorough understanding of the goals for the use of standard setting procedures and the basic components of the standard setting procedures that are commonly in use. The intent is to show how the structure and processes used for setting standards of proficiency on educational tests can be used to support a validity argument for the interpretations that are commonly made about examinee performance relative to standards located on reporting score scales.
The training session is not designed to train participants on a specific standard setting procedure. Examples will be provided of common standard setting procedures with a focus on how to evaluate whether those standard setting procedures are functioning as intended. The examples will include a discussion of the typical steps in a standard setting procedure including panelist selection, training, selection of test items, standard setting tasks, feedback, and final approval of standards. Some statistical methodology will be described and demonstrated for determining the functional characteristics of a standard setting procedure.

The training session is appropriate for anyone who is interested in learning about the conceptual underpinnings of standard setting procedures.

**Applying Data Mining Methods to Detect Test Fraud | $90**

**Thursday, April 11th | 8:45 AM - 12:45 PM EDT**

Kaiwen Man, University of Alabama; Sarah Linnea Toton, Caveon Test Security; Yiqin Pan, University of Florida

This session will provide audience with systematic training on applying various data mining models using software programs: R and/or Python to detect fraud in different test formats, such as computer based, computer adaptive or multistage settings. It covers the basics of these two software programs, theories of selected unsupervised and supervised learning methods, including K-Means, Gaussian Finite Mixture, Self-Organization Mapping, K-Nearest Neighbor, Random Forest, Supported Vector Machine, Neural Network with R/Python demonstrations. Further, the advantages and disadvantages of using each software program will be discussed.

This session consists of lectures, demonstrations, and hands-on activities of running various commonly used data mining methods. It is intended for intermediate and advanced graduate students, researchers, and practitioners who are interested in learning the basics and advanced topics related to data mining methods. It is expected the audience will have some basic knowledge of R and Python programming, but not required. Attendees will bring their own laptop and download the software programs free online. It is expected that attendees will master the basics of specifying various data mining models and applying these models to detect aberrantly behaved test-takers; further, they can apply the skills to their own research and datasets.

**Cognitive Diagnosis Modeling: A General Framework Approach and Its Implementation in R | $90**

**Thursday, April 11th | 8:45 AM - 12:45 PM EDT**

Jimmy de la Torre, University of Hong Kong; Wenchao Ma, University of Alabama

The primary aim of the workshop is to provide participants with the necessary practical experience to use cognitive diagnosis models (CDMs) in applied settings. Moreover, it aims to highlight the theoretical underpinnings needed to ground the proper use of CDMs in practice.

In this workshop, participants will be introduced to a proportional reasoning (PR) assessment that was developed from scratch using a CDM paradigm. Participants will get a number of opportunities to work with PR assessment-based data. Moreover, they will learn how to use GDINA, an R package developed by
the instructors for a series of CDM analyses (e.g., model calibration, evaluation of model appropriateness at item and test levels, Q-matrix validation, differential item functioning evaluation). To ensure that participants understand the proper use of CDMs, the theoretical bases for these analyses will be discussed.

The intended audience of the workshop includes anyone interested in CDMs who has some familiarity with item response theory (IRT) and R programming language. No previous knowledge of CDM is required. By the end of the session, participants are expected to have a basic understanding of the theoretical underpinnings of CDM, as well as the capability to conduct various CDM analyses using the GDINA package.

**Demystify Amazon Web Services (AWS): Cloud Computing and Psychometric Applications | $90**

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Huijuan Meng, Vinita Talreja, Ye Ma, CJ Taylor

Cloud computing has become increasingly popular over the past two decades and has become the cornerstone for AI applications. As practitioners who handle assessment data and do various computing tasks daily, it can be helpful to explore how cloud computing technology can be leveraged to improve efficiency and provide effective solutions to existing daunting challenges. In this workshop, we will cover several AWS core services by (1) conducting machine learning models (ML)/large language models (LLMs)/Generative AI (GenAI) analysis and (2) storing results in the database on the cloud. Participants do not need to have AWS experience. Upon completion, they will be able to streamline ML/LLM analysis on cloud with storing the results in the database using AWS services. This is a heavy hands-on training and participants are strongly encouraged to bring their laptops to follow along in order to optimize the learning outcomes from this four-hour training.

**Data Cleaning for Data Sharing Using R | $90**

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Crystal Lewis, Crystal Lewis Consulting, LLC

Before sharing research study data, it should be vetted to ensure that it is interpretable, analyzable, and reliable. This half-day, virtual workshop will provide a foundational understanding of how to organize data for the purpose of data sharing.

**Learning Objectives**

- Understand how to assess a data set for 8 data quality indicators
- Be able to review a data set and apply a list of standardized data cleaning steps as needed
- Feel comfortable using R code to clean a data set
- Understand types of documentation that should be shared alongside data

**Intended Audience**

This workshop is for any education researcher who could benefit from guidance on how to take a messy raw dataset and organize it into a shareable data product.
Any prerequisites/software required

This workshop assumes you have some experience working with rectangular data, as well as a basic working knowledge of the R programming language and experience working in RStudio. This course will focus on functions in the Tidyverse so familiarity with that package will be helpful, but is not required. All participants will need a computer and have both R and RStudio installed. More information will be provided prior to the training.

**Designing Learning Progressions to Advance Equity in Assessment and Learning | $90**

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Edith Aurora Graf, Educational Testing Service; Cheryl Lizano, Southern Illinois University Edwardsville; Peter van Rijn, ETS Global

A learning progression (LP) has the potential to provide actionable information to teachers and students by providing personalized, domain-specific information about what students already know and understand. In this training session, we focus on the design, development, and validation of an LP for the concept of function, which is central to mathematics learning in algebra and beyond. It is important that an LP is validated with the students it is intended to serve. In our work, we have focused on developing and validating an LP and associated activities for students in primarily high-needs schools. During the session, participants will engage in an experiential activity and discuss how it links to the mathematical ideas in the LP, learn how LPs can be developed and applied in a way that advances equity, see examples from steps in an LP validation cycle, and consider how small group work can be interpreted through the lens of an LP. Participants will leave the session with ideas about how to approach LP design and validation, and which psychometric methods can be used for empirical verification.

**Examining MIRT Models from a Graphical Perspective using RShiny and Mathematica | $90**

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Terry Ackerman, University of Iowa; Qing Xie, ETS

The purpose of this workshop is to walk researchers through 15 different item and test analyses based upon two-dimensional IRT compensatory model item parameters estimated from assessment data. We will illustrate each analysis graphically using the software RShiny and Mathematica. Two software packages were developed for this Workshop: MIRTGraphRS and MIRTGraphM. The packages consist of a suite of RShiny and Mathematica programs that require the user to input a file containing their estimated two-dimensional compensatory item parameters and then the user can select to run one of 15 different graphics programs. All graphics can be downloaded to include in reports or articles. The MIRTGraphRS and MIRTGraphM packages are divided into five categories of graphics for Total Test, Test Information, Two-Dimensional Ability Estimation, and Differential Item Functioning. The goal of this workshop is to provide graphical tools that will yield greater insight into what tests are measuring and inform future test development.

The session will have an interactive format where the audience can use their laptops and follow the
presenters as they describe the various graphics. Attendees will receive a manual, detailing each of the graphics, and a copy of all the R Shiny and Mathematica programs that are discussed.

**K-12 Assessment & Accountability Work in Practice | $90**

Thursday, April 11th | 8:45 AM – 12:45 PM EDT

Andre Rupp, Center for Assessment / National Center for the Improvement of Educational Assessment; Laura Pinsonneault, NCIEA; William A. Lorie, Center for Assessment; Juan Manuel D'Brot, NCIEA

Have you ever wondered what kind of advanced methodological work is relevant in the area of K-12 assessment and accountability? Have you been curious about how interdisciplinary expertise from across diverse educational areas is relevant to evidence-based decision-making in this space? If so, then this workshop could be for you!

In this half-day workshop, we discuss conceptual frameworks and applied use cases that illustrate common methodological problems in the area of K-12 assessment and accountability based on published work. Throughout we offer our own critical reflections and reflect with the participants on the implications of this work for postsecondary training programs as well as on-the-job training.

Methodologically, we cover a range of issues that touch upon the relevance of data engineering and governance issues, simulation studies for applied research, and evidentiary reasoning for federal audits. We also discuss foundational tools for the development of theories of action, balanced assessment systems, and targeted program evaluation with special considerations for how to address equity issues programmatically. We provide a library with relevant additional readings via a GoogleDrive as well as through Zotero, a free digital resource management program. In addition, we connect with participants through pre- and post-workshop surveys and outreach.

**Organizational Leadership for Measurement Experts | $90**

Thursday, April 11th | 8:45 AM - 12:45 PM EDT

Ye Tong, yetong@nbme.org; Michael C. Rodriguez, University of Minnesota; Jennifer Dunn, Pearson Assessments; Richard Patz, UC Berkeley

Participants in this training course will learn to identify the skills needed to function effectively as an organizational leader, learn from the real-world experience of measurement professionals who have been effective in leadership roles, and have the opportunity to reflect upon and explore their own interests and skills. Open to all but specifically intended to address the interests of early and mid-career professionals who are engaged in or contemplating organizational leadership roles, individuals from all backgrounds, including those from groups traditionally under-represented in leadership roles, are highly encouraged to participate. The facilitators have both rigorous scientific training and extensive organizational leadership experience. Leveraging NCME's Foundational Competencies in Educational Measurement, dimensions of effective leadership are presented and discussed broadly, informed by real-world experience from within testing organizations and higher education contexts. The training session will be interactive, affording opportunities for self-reflection, small-group activities, and large-group discussions.
Bayesian Latent Variable Modeling in Education Research | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Edgar Merkle, University of Missouri

The goal of the session is to provide background on Bayesian latent variable modeling and to illustrate the models using education data. We will consider factor analysis models, item response models, structural equation models, and two-level variations, with distinctions between the models becoming especially blurry under a Bayesian viewpoint. Specific topics include theoretical background, advantages and disadvantages of adopting a Bayesian modeling approach, prior and posterior checking, model estimation, and model extension. The topics will be illustrated via case studies in R, especially focusing on the blavaan package paired with Stan. Session attendees are assumed to have some knowledge of item response and/or structural equation models, though not necessarily in depth.

Building Monte Carlo simulations in R for Measurement, Causal Inference, and Beyond! | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Luke Weisman Miratrix, Harvard Graduate School of Education

In this course we will learn how to write Monte Carlo simulations in R. Monte Carlo simulations are an essential tool of inquiry, useful both for small-scale investigations and for formal methodological research. Simulation can assess, for example, how much re-testing biases a regression discontinuity design due to undermining the measurement properties of a test score. Our focus is on the best practices of simulation design. Overall, we will show how a specific simulation framework allows for rapid exploration of the impact of different design choices, measurement qualities, and data concerns, and show how simulation can answer questions that are hard to answer using direct computation. For example, available algebraic formulas are often based on asymptotic approximations, which might not “kick in” if sample sizes are moderate; simulation can provide an answer.

In this session we will work through two simulation case studies, showcasing a modular programming approach for good design. Code, along with an on-line open-source textbook, will be provided and demonstrated. Students should bring laptops, with R and RStudio installed, to follow along. By the end, students will be able to adapt provided code to their own purposes moving forward. Some familiarity with R is assumed.

Comprehensive Statistical Model Evaluation: Traditional, Bayesian, and Information-theoretic Methods and User-friendly Software | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Wes Bonifay, University of Missouri (MU); Sonja D. Winter, MU; Hanamori F. Skoblow, MU; Brian T. Keller, MU
Statistical models play a central role in scientific analysis, inference, and decision-making, so it is imperative that researchers diligently and thoroughly evaluate their models before disseminating them. This training session offers an immersive exploration of three perspectives on statistical model evaluation. Attendees will gain a theoretical and methodological understanding of (1) traditional goodness-of-fit testing and bootstrapping procedures, (2) Bayesian prior and posterior predictive model checking, and (3) information-theoretic techniques that adhere to the principle of minimum description length. This discussion will culminate in a simple framework that integrates all three perspectives. Session leaders will then demonstrate a user-friendly Shiny software application that allows users to upload data, specify a statistical model, select any or all of the methods within the framework, and generate a customized model evaluation report. Attendees should bring their own laptops, with R and RStudio installed prior to the session.

The intended audience comprises researchers who use statistical models of any form. The methods covered in this session have been applied to item response theory, factor analysis, and structural equation models, and could be extended to other modeling frameworks. There are no prerequisites for this session, other than general familiarity with the practice of statistical modeling.

**Computerized Multistage Testing: Theory, Practical Issues, and Solutions (Book by Routledge) | $90**

Thursday, April 11th | 1:00 – 5:00 PM EDT

Duanli Yan, ETS; Alina A. von Davier, Duolingo; Kyung (Chris) T. Han, GMAC

This course provides a general overview of a computerized multistage test (MST) design and its important concepts and processes in the age of AI. The MST design is described alongside the considerations for test development, psychometrics, and engineering. We will also discuss how it differs from other test designs, such as a linear test and a computer adaptive test (CAT), and how to design and conduct simulations for an optimal test.

**Diagnostic Classification Modeling with R and Stan | $90**

Thursday, April 11th | 1:00 – 5:00 PM EDT

W. Jake Thompson, University of Kansas

Diagnostic classification models (DCMs; also known as cognitive diagnostic models [CDMs]) have gained interest in recent years due to their ability to provide fine-grained actionable feedback while keeping test lengths short. In this workshop, we will cover how to easily estimate and evaluate DCMs with R and Stan. The workshop will include hands-on examples of defining a DCM, estimating the model, and evaluating the fit (e.g., test- and item-level fit, classification accuracy and consistency, etc.).

The goal of the workshop is to enable participants to implement DCMs in their own work, and thus is intended for anyone who uses, or would like to use, DCMs for applied or research uses (e.g., psychometricians, faculty, applied researchers, graduate students). Although not necessary, prior experience with R will be helpful. All workshop materials, including slides, examples, and solutions will be available on a workshop website. Participants should have access to a laptop they can bring to the workshop in order to follow along with the examples. Instructions for installing any necessary software will be provided.
Educational Testing and Psychometrics with R Package dexter | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Ivailo Partchev, CITO

Among the 250 core and regular packages listed in the psychometric view of CRAN, dexter stands out by offering a comprehensive system for the management and analysis of test data. This includes a solid SQL data base created automatically, numerous and advanced diagnostic techniques for items and tests, and efficient support for ability estimation (both standard and plausible value based). dexter can calibrate connected multi-booklet designs of great complexity, and analyze large data sets very fast. It includes many innovative techniques for DIF detection, equating, profile analysis, and more. An easy to use graphical interface is provided by the companion package dextergui, while another package, dexterMST, supports a version of multi-stage testing particularly attractive in the context of high stakes exams. dexter is appropriate for production purposes and actually used in production on a daily basis. It was published in the hope of making IRT more appropriate, attractive, and accessible for the many countries and testing establishment around the world that still rely exclusively on classical test theory and traditional equating methods.

Longitudinal Diagnostic Classification Models: Theory and Applications | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Matthew James Madison, University of Georgia

Diagnostic classification models (DCMs) are psychometrics tools that focus on providing actionable feedback in the form of examinee attribute classifications. Longitudinal DCMs have been developed and applied as psychometric tools for analyzing diagnostic assessments administered over multiple occasions. Different from traditional psychometric frameworks for growth, which typically provide continuous and norm-referenced growth estimates, longitudinal DCMs provide categorical growth estimates with criterion-referenced interpretations. This workshop focuses on longitudinal DCMs and their application as the psychometric foundation for categorical growth models. After completing this workshop, participants will understand the structure of longitudinal DCMs, be able to estimate longitudinal DCMs using a newly developed R package, and interpret software output.

This session is appropriate for graduate students, researchers, and practitioners at the emerging or experienced level. Participants are expected to have a basic knowledge of DCMs and psychometrics to enroll. This session presents both conceptual and technical content and also provides hands-on experience for participants to apply what they learn. Content will mostly be delivered through lecture, and content will be reinforced using hands-on activities. Instructor will encourage audience participation through questions and allow time for discussions among participants and the instructor.

Optimal Test Design Approach to Fixed and Adaptive Test Construction using R | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Seung W. Choi, University of Texas at Austin; Sangdon Lim, UTA
Many testing programs use fixed test forms and computerized adaptive testing (CAT) forms interchangeably, but conventional CAT selects items through computer algorithms while fixed forms are created using iterative review processes. The optimal test design framework provides an integrated solution for creating test forms in various configurations and offers a simplified workflow for meeting complex blueprint requirements. This workshop covers the principles of the optimal test design framework and their applications in fixed and adaptive test construction, along with practical examples and an R package for creating and evaluating various fixed and adaptive test formats.

Professional Training for Graduate Students in Measurement | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Deborah J. Harris, The University of Iowa; Nathan Wall, EMETRIC; Yi-Fang Wu, Cambium Assessment, Inc.

This training session will address practical topics for graduate students in measurement to find a job and start a career. First, what to do now while they are still in school to best prepare for a job, which includes the types of training employers look for and how to obtain it (classes, workshops, online training, etc.), how to find a topic and complete a dissertation, how to maximize experiences with networking, internships, social media, and volunteering. Second, how to locate and interview for a job, which includes finding open positions and the application process, including tailoring cover letters, references, and resumes. Third, what to expect in the interview process, including online and in-person interviews, job talks, questions to ask, and negotiating an offer. Last, starting a career, adjusting to the work environment, evaluations, and people (clients, students, co-workers, bosses, mentors, etc.), establishing a career path, searching for grants/funds/awards opportunities, work-life balance, dealing with a bad fit, and staying current. The session addresses working-from-home, publishing, professional associations and service, layoffs, and the wide variety of jobs available. The session is interactive, geared to addressing the attendees’ particular interests during the session, and providing resource material on all topics as a takeaway.

Vertical Scaling: Hands-on Practice and Evaluation of IRT Linear and Non-linear Methods | $90

Thursday, April 11th | 1:00 – 5:00 PM EDT

Hyeon-Joo Oh, Riverside Insights; Tim Moses, College Board; Hanwook Yoo, Educational Testing Service

This training demonstrates the empirical application of vertical scaling approaches developed for newly launched large-scale assessments. Vertical scales are a current interest in K-12 testing programs for supporting the measuring and tracking of growth across grade levels. This training session focuses on the following topics: (1) conceptual and technical background of vertical scaling, including the definition of student growth, data collection design, and scaling methods, (2) step-by-step guides to conduct item response theory (IRT) vertical scaling with linear and non-linear methods of converting student ability estimates to scaled scores, (3) a hands-on exercise related to evaluating scaling outcomes and comparing scaled scores, and (4) practical issues and future research topics on vertical scaling. Through the training session, we hope participants will achieve empirical vertical scaling work experience and an understanding of the challenges and limitations related to vertical scaling that psychometricians face in new test development. The training is targeted toward advanced graduate students or new
measurement professionals who want to practice vertical scaling. We recommend that participants have some background knowledge in IRT-based equating and scaling, but it is not required. Attendees should bring their own laptops to conduct the hands-on activities.