**EvidenceStream: Streaming Evidence Identification and Accumulation for Game- and Simulation-based Assessment**

**Overview**

Game- and simulation-based assessments (including instructional systems with embedded assessments) are opening up new educational opportunities, especially in STEM fields. However, these systems can log tens of thousands of events (user actions) per user, and users typically interact with the systems across days or weeks of time. Evidence-centered design (ECD) has proved successful in defining which events provide evidence of the STEM competencies educators wish to assess in prior NSF-supported research, such as the game *Physics Playground*. Building scalable learning tools requires a software framework for streaming analysis of the potentially millions of events into inferences about a handful of competencies. The proposed *EvidenceStream* is such a framework, along with tools for helping developers translate pedagogical knowledge about the competencies into scoring algorithms.

# Intellectual Merit

# In ECD, steps in the scoring process are called *evidence identification* and *accumulation*. Evidence identification determines the values of key observable variables within a *scoring window* (a game level or a single task in a simulator). Evidence accumulation summarizes across scoring windows to update beliefs about the competencies. The ECD framework has been used in a number of existing assessment systems, including the NSF-sponsored game *Physics Playground*. The prototype implementation used in *Physics Playground* uncovered two major issues: 1) evidence identification and accumulation needs to be more efficient, taking advantage of parallel processing; and 2) better training is needed as the description and debugging of scoring models is a complex operation.

The proposed research would build *EvidenceStream*, an open-source scoring toolkit which will allow designers to build streaming scoring services for game- and simulation-based assessments. *EvidenceStream* is a redesign of the existing prototype scoring tools explicitly based on a streaming architecture with three stages, each of which refines the data: *evidence preparation* will filter unused events and add labels to key events, *evidence identification* will condense multiple events within a scoring window into key *observables*, and *evidence accumulation* will use observables to update beliefs about competency variables. There is already a robust Bayesian network toolkit for the evidence accumulation. The evidence allocation would be based on a rule-based system designed for *Physics Playground*, but redesigned to both operate on parallel computing and be easier to code. There are a number of statistical and machine learning models that can be used to label events in the evidence classification stage. In this manner, *EvidenceStream* will provide educational researchers with an efficient tool for extrapolating viable data from game-based assessments when investigating questions of student competency in STEM fields.

# Broader Impacts

# Releasing *EvidenceStream* as free open-source toolkit will facilitate the development of game and simulation-based assessment. In particular, the evidence identification (classification and allocation) stages have few available tools, yet play a vital role in data reduction. Projects that could immediately benefit from this include *Physics Playground* (Florida State University; Newtonian physics) and *e-Rebuild* (Florida State University; middle school math), *Zoombinis* (TERC; computational thinking) and *SimScientist* (WestEd). However, the broader impact will come from the support of new assessments, especially in STEM fields where the ability of students to interact with simulations is critical to learning science and engineering practices. The tools will be available researchers and developers in STEM and other educational fields through open-source, royalty-free releases. To maximize the impact, the project will develop free online training material for building scoring models.

The same architecture is applicable to types of data that are not generated by games and simulations. In particular, audio and video data also come in streams and can be scored through the same sequence of processes: *evidence preparation*, *evidence identification*, and *evidence accumulation*. In particular, the evidence allocation allows the systems to respond to sequences of events or events in contexts. This will be useful in a wide variety of educational and training contexts in STEM fields and beyond.