Title: Bridging the Lab and the Field with *In Vivo* Experimentation
Abstract Body

Background / Context:
Some large-scale randomized field trials in education have demonstrated significant treatment effects, but many have not (see www.whatworks.ed.gov). Given that a large fraction of randomized field trials yield null results, how can we do better empirical and theoretical science leading up to such trials? We need foundational theory development that is based on more focused experimentation that is both reliable and relevant. Even when field trials are successful, the lessons are typically limited to the particular curriculum or instructional treatment under question (and often only relative to the comparison curriculum or condition used in the study). We learn little about what features of a successful treatment might be successfully incorporated in instructional designs for other courses and student populations.

Purpose / Objective / Research Question / Focus of Study:
To produce more cumulative educational research (NRC, 2002), we need a transition between laboratory studies and field trials. A lot of attention has been paid to internal validity both in laboratory studies and field trials. Laboratory research does not pay significant attention to ecological validity, that is, whether the learning principles under examination in rarefied laboratory conditions will stand up in the context of real classrooms, student motivations and academic content. Neither approach has attended much to external validity, that is, how broadly do the approaches or principles under study generalize to other situations and settings. Our purpose is to discuss a research methodology, called in vivo experimentation that balances internal, ecological, and external validity to produce educational theory that is reliable and relevant to practice. As illustrated in Table 1, in vivo experimentation, sits within a space of methodologies within the educational and learning sciences. Differences are described below.

Setting:
In vivo experiments are run as part of functioning courses and are typically embedded within educational technology components (e.g., intelligent tutoring systems) of those courses. Most often students are using the technology in a classroom or computer lab as part of a regularly scheduled class session. In some cases, students may be on their own working at home, in their dorm, or on a portable laptop.

Population / Participants / Subjects:
Participants in in vivo experiments are students taking a course for credit, typically in the junior high to early college range. Typical sample sizes are 50 to 300 students from two to ten sections of a course.
**Intervention / Program / Practice:**
To facilitate a more reliable transition from laboratory research to field trials, the Pittsburgh Science of Learning Center’s (PSLC) LearnLab has been supporting researchers in conducting “in vivo experiments” and associated theory development around the instructional principles tested in those experiments (see learnlab.org). An *in vivo* experiment is a laboratory-style principle-testing experiment conducted in the classroom. The conditions manipulate a small but crucial, well-defined instructional variable, as opposed to a whole curriculum or multi-faceted approach. *In vivo* experimentation is differentiated from lab experimentation in its emphasis on ecological validity (real students, content, setting, motivations) and from many randomized field trials in its emphasis on varying a single theoretical principle (not a whole curriculum, text, or extended practice). Table 1 summarizes key differences from other methodologies including the “design experiment” approach (e.g., Barab & Squire, 2004) used by many learning science researchers. *In vivo* experiments tend to have a cost and duration (a month) that is intermediate between shorter (an hour) and lower cost lab experiments and longer (a year) and higher cost design experiments and randomized field trials.

*In vivo* experiments are run in schools or colleges with which PSLC has standing agreements to work with instructors at those sites to implement studies (usually running for a few weeks) once every semester in one of the courses that they use. Each “LearnLab course” has a committee of researchers and instructors that meets monthly to plan studies, insure quality, and help interpret results. The PSLC LearnLab has course committees and sites for courses in Algebra, Geometry, Middle School Math, Chemistry, Physics, Statistics, English as Second Language, and Chinese. All of these courses blend uses of educational technology (intelligent tutoring systems, virtual labels, on-line interactive readings, discussion forums) with face-to-face instruction and discussion.

**Research Design:**
This paper is a summary reports of *in vivo* experiments, smaller-scale randomized field trials that vary one theoretical principle at a time. These experiments typically focus on a single curriculum unit and last a few weeks. They involve two or more of the following conditions: 1) an *ecological control*, that is, current course practices in the target unit, 2) a *treatment* that varies a single instructional factor relative to the ecological control or a tight control (see #3) in order to test a specific theoretically-based instructional principle, or 3) a *tight control* that is one factor different from the treatment and is used in cases where the treatment and ecological control differ on multiple factors. PSLC provides technical support and tools like the Cognitive Tutor Authoring Tools (see ctat.pact.cs.cmu.edu) to aid researchers in implementing treatments and tight controls (often by modifying the existing technology used in the ecological control).

Random assignment at the student level is achieved when possible, particularly when the varying instructional factor involves a more subtle change to a technology component that is not clearly
apparent to students (e.g., a difference in the algorithm in an intelligent tutor that selects the next activity based on its student model). In cases where larger, more visible changes are made (e.g., students working pairs rather than alone), random assignment is done at the class section level.

**Data Collection and Analysis:**

The main source of instructional effectiveness data form *in vivo* experiments is pre- and post-assessments, often performed on paper, but sometimes given on-line in addition or instead. Post-assessments typically include not only an immediate test of the direct objectives, but one or more measures of robust learning including long-term retention, procedural and conceptual transfer, acceleration of future learning, or increased desire for future learning.

Because *in vivo* experiments are often run with a technology component, detailed on-line keystroke data is typically collected. Such data is temporally dense with a typical rate being an observation of a student action about six times a minute. Such on-line data, as well as raw summaries of paper based pre- and post-test data, are available for secondary analysis at PSLC’s DataShop (see pslcdatashop.org), an open repository and tool set for educational data mining, which houses over 150 student learning datasets.

Typical analysis methods include ANOVA, MANCOVA, or some instance of a mixed effects generalized linear regression. On-line data is typically analyzing using logistic regression and growth curve models of learning curves with student success rate (e.g., correct on first attempt at step without a hint request) and number of practice opportunities as the growth variable. The DataShop (pslcdatashop.org) automatically runs such learning curve models on standard data sets and facilitates comparison of alternative cognitive diagnosis models of student performance and learning.

**Findings / Results:**

PSLC has supported researchers in running over 100 *in vivo* experiments in live courses in high school or college mathematics, science, or second language. These studies have been exploring a variety of instructional principles many of which are inspired by basic cognitive research. A number of these principles overlap with recommendations from a “practice guide” produced by the Department of Education (Paschler et al., 2007), including spacing of practice, extensive use of worked examples, and prompting students for self-explanations. This guide honestly and accurately indicated the moderate strength of evidence behind these recommendations. PSLC research has not only been contributing to the strength of evidence for these recommendations, but exploring their theoretical and empirical scope. Results of experiments in language courses have demonstrated for instance that wide spacing of
practice is not always the most effective and efficient approach to enhancing student learning (e.g., Pavlik, 2006; Pavlik et al, 2007).

Multiple experiments across a number of math and science domains (algebra, geometry, physics, and chemistry) have provided further external validity support to the value of extensive use of worked examples and of prompting students for self-explanation (e.g., Aleven & Koedinger, 2002; Hausmann & VanLehn, 2007; McLaren, Lim & Koedinger, 2008; Salden et al., 2008). At the same time, studies in a second language course suggest boundaries on the scope of the self-explanation effect (Wylie, Koedinger, Mitamura, 2009). While having students study examples and explicate domain rules in their own words appears to significantly enhance transfer of learning in math and science, it appears that such rule explication may slow down learners of a second language.

Conclusions:

In vivo experimentation is now a well-established methodology that balances advantages of laboratory studies and field trials toward producing results that are reliable and relevant to practice. By contributing to cumulative development of theory, courses, and instructional practices, it may yield better preparation for large-scale randomized field trials that produce large effects. For instance, this process contributed to the design of on-line interactive instructional materials for statistics (see oli.web.cmu.edu/openlearning) that are used as part of a “blended” course with face-to-face meetings as well as on-line learning. A randomized field trial comparing the blended course to a traditional course demonstrated a huge improvement in student learning (Lovett, Meyer, Thille, 2008). Students using the on-line course learned more (18% vs. 3% improvement in the standardized CAOS test of statistical reasoning concepts) in only half the time (a half semester rather than a full semester and no extra out of class time)!

Critical to the in vivo experimentation methodology is the social and technical infrastructures needed to support it. On the social side, PSLC has developed practices and procedures, like LearnLab course committees for researcher-instructor co-design and study vetting, memorandums of understanding with school sites to document shared understanding (e.g., an experiment per class per semester limit), and a full time site manager to maintain relationships with school sites. On the technical side, PSLC provides authoring tools and technical support for developing technology components and data collection and analysis tools for the detailed keystroke-level learning process data. This infrastructure is both a strength and a limitation and future efforts are needed to broaden the community of users that can take advantage.
Appendix A. References


