STUDENT LEARNING AND ANALYTICS AT MICHIGAN

February 8, 2013:

Learning Analytics, Learning Metrics and Learning Science

David Niemi, Vice President of Measurement and Evaluation, Kaplan, Inc.

STUDENT LEARNING AND ANALYTICS AT MICHIGAN

www.crlt.umich.edu/slam

Learning Measures, Learning Analytics, and Learning Science

David Niemi
Kaplan

University of Michigan, Ann Arbor
February, 2013

Starting points

- Learning analytics should be informed by learning science
  - Little evidence to suggest that key findings of cognitive and learning science have had any significant impact on postsecondary curricula, teaching, training, instructional design, or assessment
- Good learning analytics require valid measures
- Explosive growth of online education provides new opportunities for analytics and research to improve learning (e.g., Shadish & Cook, 2009)

"Tonight, we're going to let the statistics speak for themselves."

What can we do with valid learning measures?

- Improve teaching and learning
  - Formative improvement of teaching and learning
  - Drive adaptive or personalized instruction
  - Identify students at risk of failing or withdrawing
  - Diagnose skills, misconceptions, deficiencies
  - Measure individual growth and provide feedback for students and instructors
  - Measure effectiveness of innovative approaches
- Evaluate curricula, programs, instruction, instructional design
  - Most/least effective strategies
  - Areas of relative strength/ deficiency
  - For sub-groups and all students
- Placement
- Promotion, graduation, certification decisions
- Career planning

Measuring learning

- Learning: change in knowledge and skills over time
- Measurement: putting some quantity on a scale
  - E. g., grades?
- To measure learning requires valid (accurate and reliable) and equivalent measures at two or more points in time--for the same learners

Kaplan basic learning metrics

- % of course or lesson objectives student masters
- % of students achieving mastery of course or lesson objectives
  - Requirements:
    - Well-defined learning objectives or outcomes for programs and courses
    - Valid assessments for each objective
    - “Mastery score” for each objective; e .g., 80% on relevant assessments
      - Allows you to calculate % of objectives that each student has mastered
    - Mastery target for each course or program; e. g., “Students should master 80% of objectives”
      - Allows you to calculate the % of students achieving the target

Guide to designing learning assessments

Matching knowledge components and assessments

Knowledge Component		Assessment			
		Remember	**Proxy for Remember**	**Use**	**Proxy for Use**
Procedure		Recall when to use	Reorder steps	Decide when to use	Critique performance or output of actions and decisions
		Recall action and decision steps	Recall next or missing steps	Perform the steps (actions and decisions)	
Supportive Conceptual Knowledge	Fact	Recall fact	Recognize fact when presented with distractors	Recall fact in task context	
	Concepts	List defining attributes verbally or in writing	Recognize defining attributes when presented with distractors	Identify or generate examples and non-examples	Critique someone else's identification or generation of examples
	Process/	Recall phases, events, and causes	Recognize phases, events, and causes	Identify causes of faults in a process	Critique someone else's description of causes or prediction of events in a process
	System		Recall missing phases, events, and causes	Predict events in a process	
	Principle (cause and effect relationship)	Recall the principle	Recognize the principle	Decide if a principle applies	Critique someone else's application of the principle to solve a problem, explain a phenomenon or make a decision
			Recall missing elements of the principle	Predict an effect	
				Apply the principle to solve a problem, explain a phenomenon or make a decision	
Knowledge Integration				Opportunities (including instructions, templates, rubrics) to self-explain, discuss, present, describe or select reasoning about interconnections among knowledge components, for example the principle(s) that justify the application of a procedure Explain the interconnections among conceptual knowledge components, or the conceptual foundation of procedures, or the procedural implementation of conceptual knowledge components	
Knowledge Transfer				Demonstrate ability to use knowledge components in multiple and varied contexts	
				Explain how to use knowledge components in other contexts	

Reliability studies of Course Level Assessment (CLA) and General Education Literacy (GEL)

Description of CLAs and GELs

- 4-6 CLAS in each course, tied to major course learning objectives
  - Example of CLA objective:

    "Describe typical neurobiological and behavioral responses to stress and their implications for physical and mental functioning."
- 1-2 GELs in most courses, tied to program learning objectives
  - Example of GEL objective

    "Demonstrate college-level communication through the composition of original materials in Standard American English. "
- Examples of CLA and GEL tasks:
  - Projects, essays, online discussion boards
- Office of Institutional Effectiveness wrote manual on designing outcomes, tasks and rubrics

- Conducted study within a domain specific assessment framework (Course Level Assessments or CLAs) developed by an online university
  - Assessments implemented in 1000 online courses, ranging from arts and sciences to business, education, nursing, criminal justice and legal studies, information technology, finance and other commonly taught post-secondary subjects
  - University has an extensive database of information on students and their performance, including measures of learning, student satisfaction, postgraduate success, etc.
- Also studied writing skills across courses (General Education Literacies or GELs)

CLA and GEL studies

Hypotheses

- Student's CLA scores should remain steady over time.
- GEL 1.1 (writing) scores should be stable across courses taken simultaneously – in other words, scores should be consistent for a single student who is rated twice in the same term by 2 different instructors teaching 2 different courses.
- Student's GEL scores should increase or stay the same over time.

CLA and GEL Studies

Results

- The model identified faculty scoring as the greatest source of variation for both CLA as well as GEL scores.
- After numerous iterations were run for different tracks and degree programs, the model revealed greatest variation for faculty scoring in the Associate's and Bachelor's programs and the least – for Master's program faculty.
- For the fixed effects, the only statistically significant predictor was full-time/part-time status of the faculty with the full-time faculty outperforming their part-time counterparts.

CLA and GEL Studies

Results (continued)

- After additional analyses were performed on just GEL 1.1 scores to identify the variability between instructors teaching the same student in the same term, the results showed significant variation due to the instructors' variability.
- The model identified the difference amonh faculty as the greatest source of variation for the student GEL scores (apart from the unexplained (residual) variation).

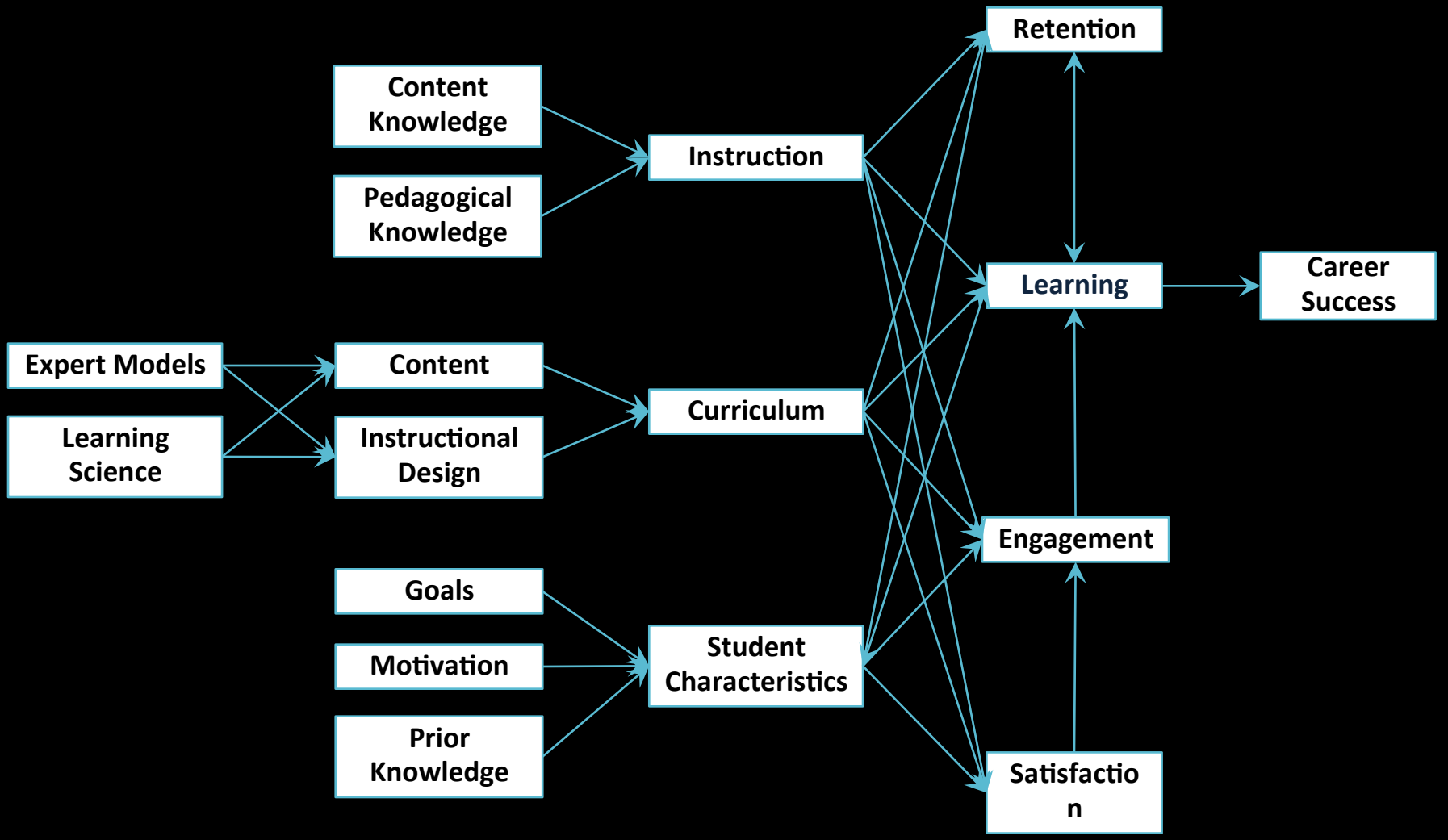
Other CLA reliability studies

- Results of rescoring studies in two different courses suggest that
  - Faculty score their own students higher than other faculty do
  - Faculty do not agree on the scores for the same papers
- Need more replications
- Test rubric simplification, training strategies
- Try Western Governors' scoring approach?
- Try automated scoring?

Other metrics

- Changes in scale scores (learning)
- Performance in subsequent courses (learning)
- Engagement
- Retention
- Satisfaction
- Career or future academic success

Key outcomes and the factors that influence them



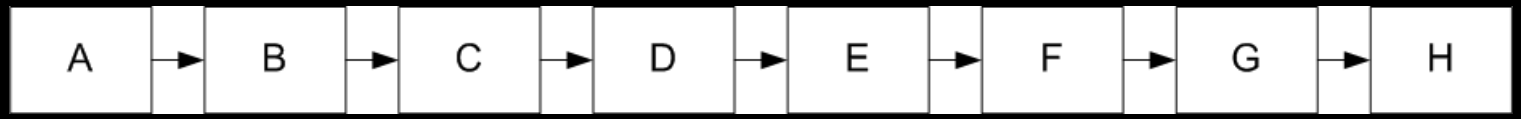
Using assessment data to build a cognitively-sensitive assessment system

- Compare learning sequences or paths
  - defined by experts
  - indicated by measurement data
- Goal
  - create learning paths that are supported by measurement data and accepted by subject matter experts

- Process
  - determine which of the expert-identified topic sequences were supported by total population statistics and which were not
  - where the expert sequence was not supported, relationships were modified through expert judgment
  - incrementally discover a sequence (precedence structure) supported both by experts and by the measurement data from classical item and test analysis (p-value and point biserial)
- Side benefits
  - improved item alignment
  - better grouping of skills within curriculum

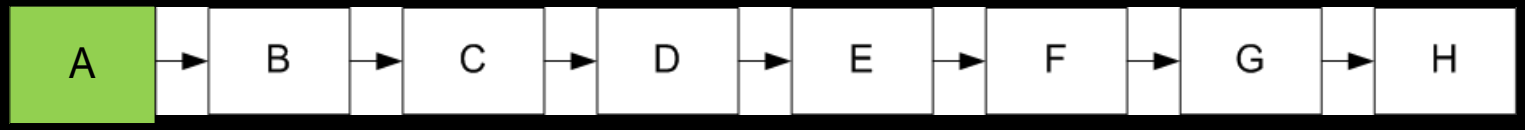
Background: Linear Paths

- Learning Sequence (e.g. Syllabus)



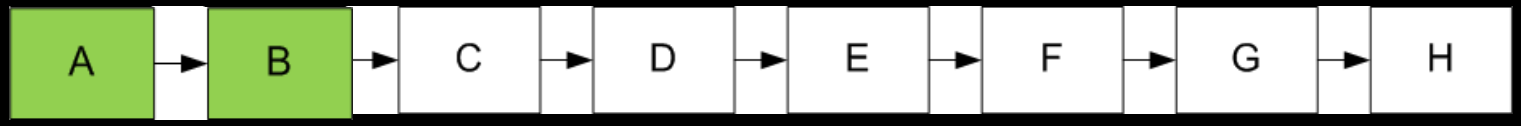
Background: Linear Paths

- Testing/Measurement Pattern



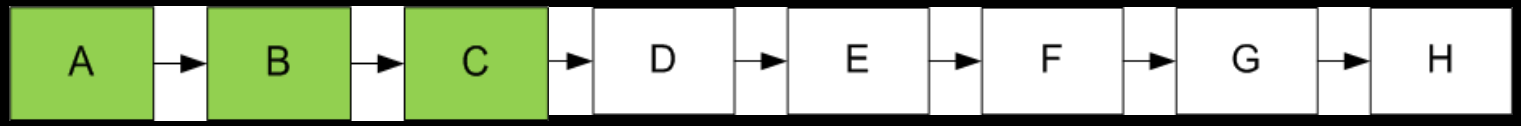
Background: Linear Paths

- Testing/Measurement Pattern



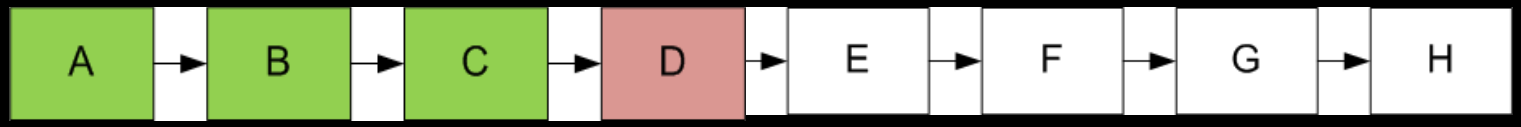
Background: Linear Paths

- Testing/Measurement Pattern



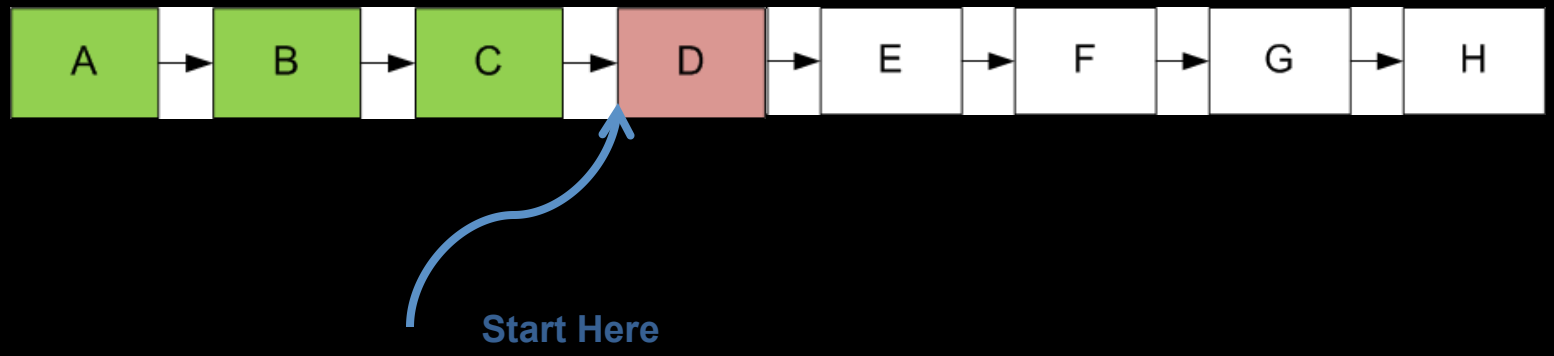
Background: Linear Paths

- Testing/Measurement Pattern



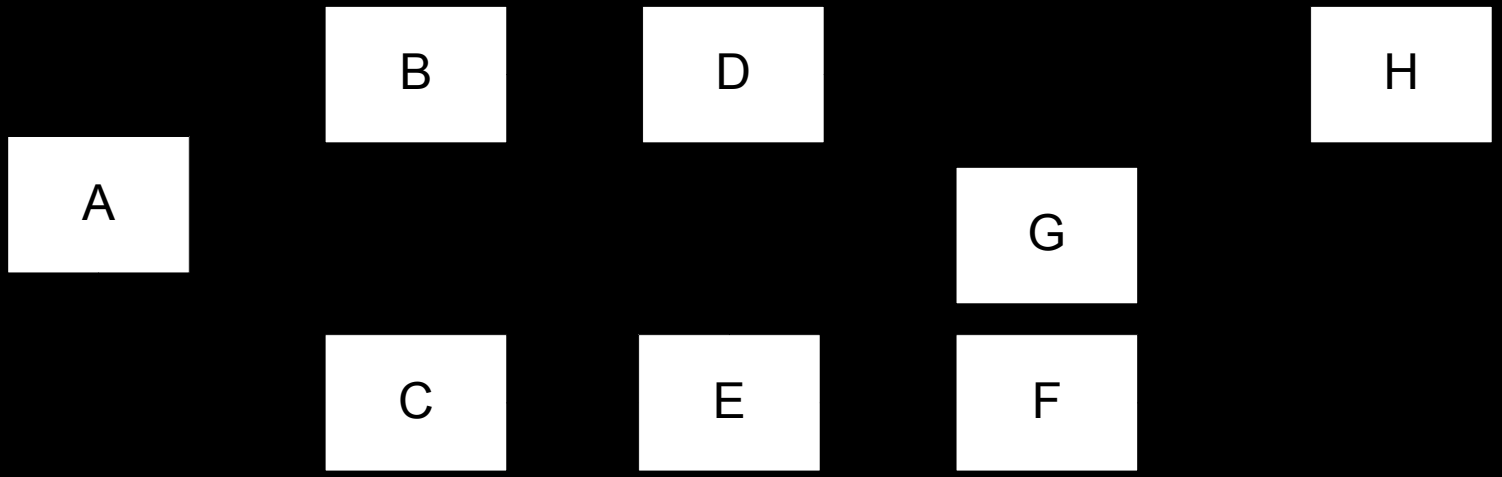
Background: Linear Paths

- Instructional Recommendation



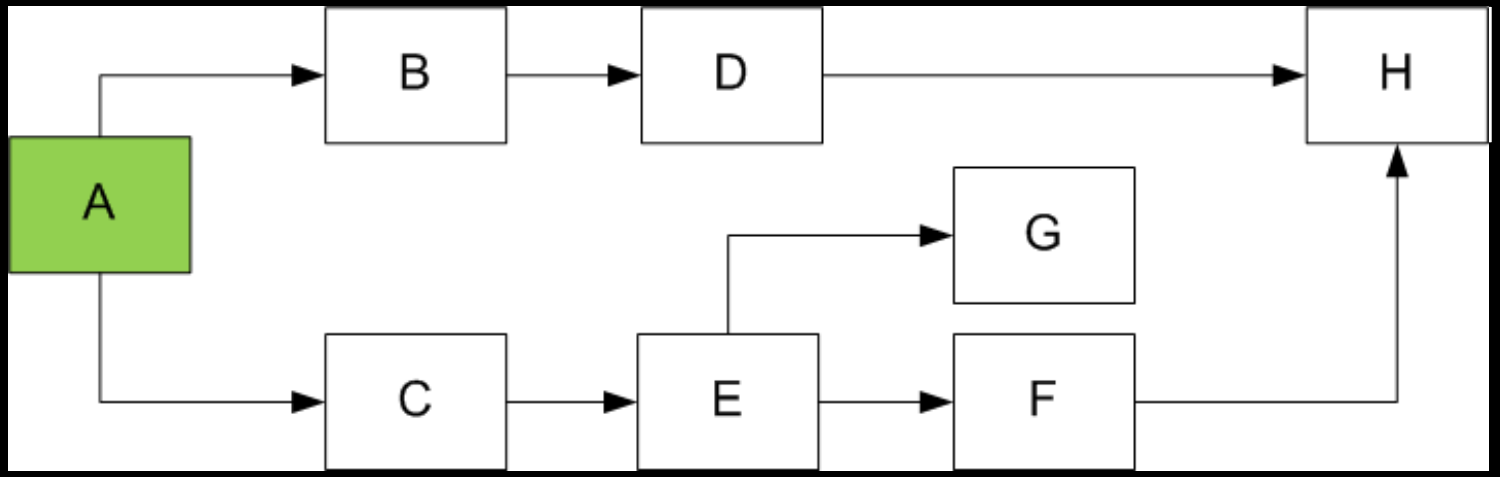
Background: Multipath

- Learning Progression or Map



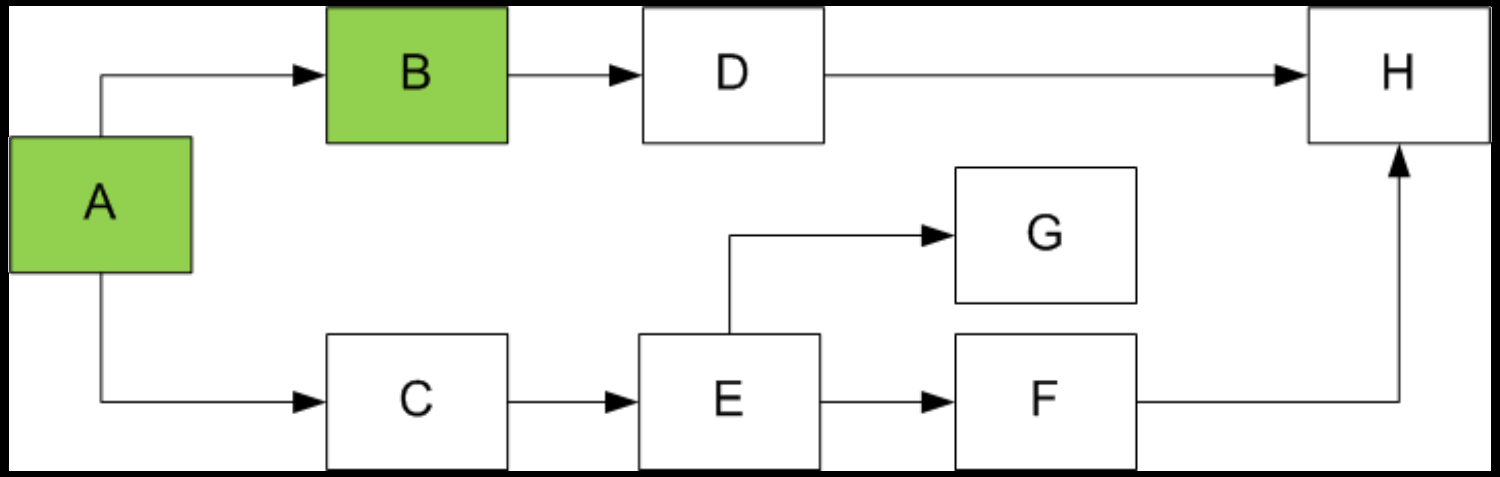
Background: Multipath

- Measurement Pattern



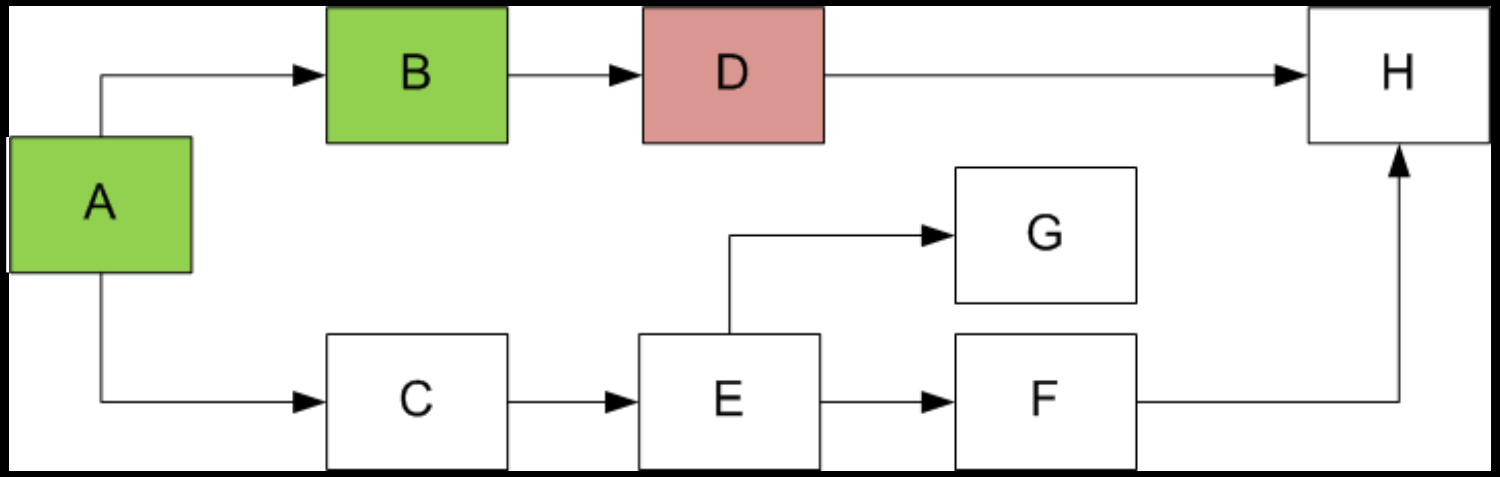
Background: Multipath

- Measurement Pattern



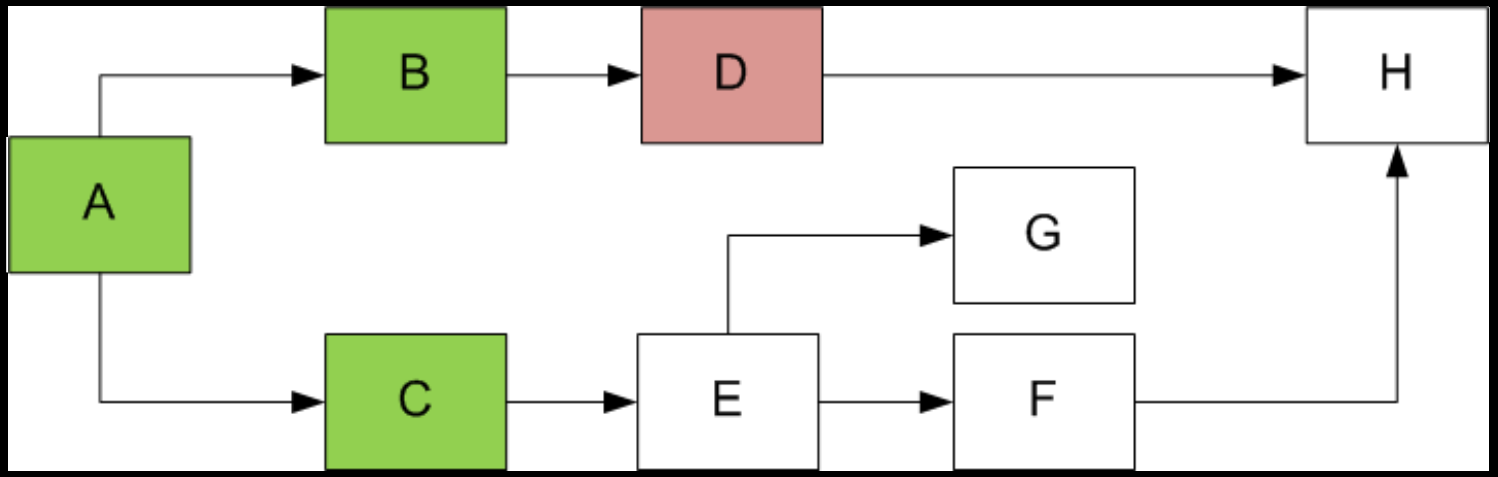
Background: Multipath

- Measurement Pattern



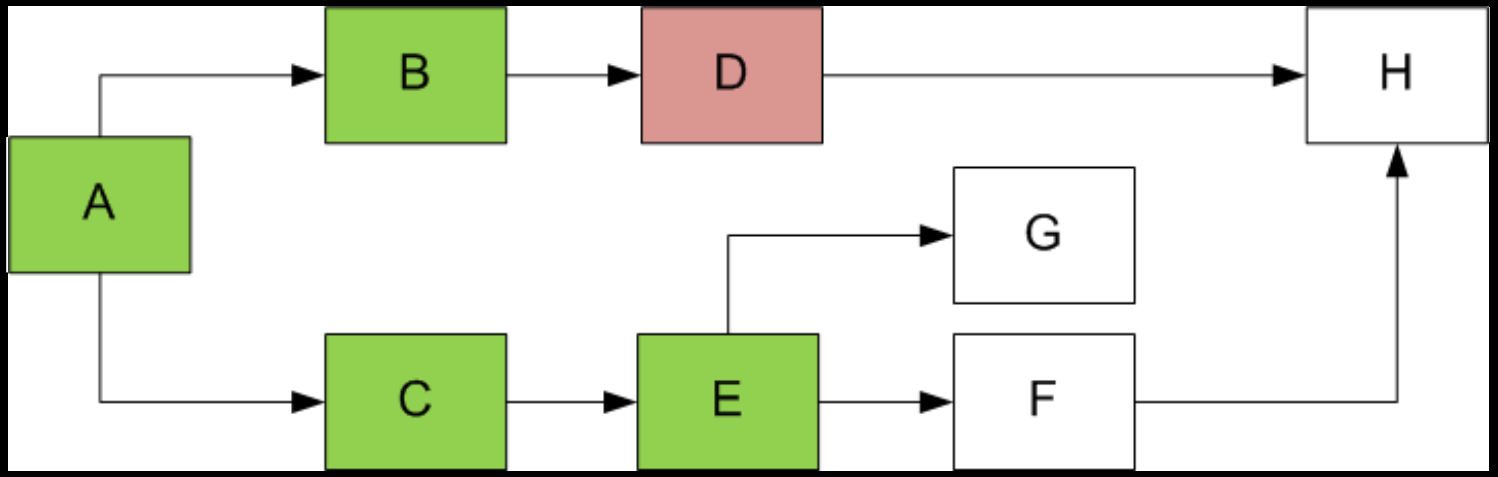
Background: Multipath

- Measurement Pattern



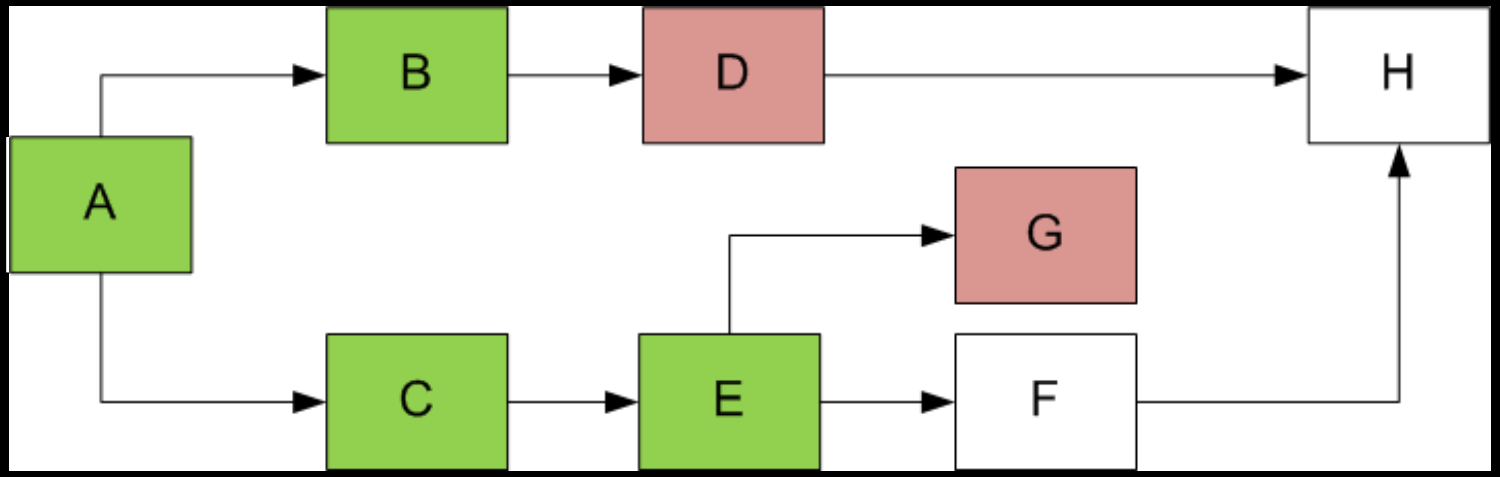
Background: Multipath

- Measurement Pattern



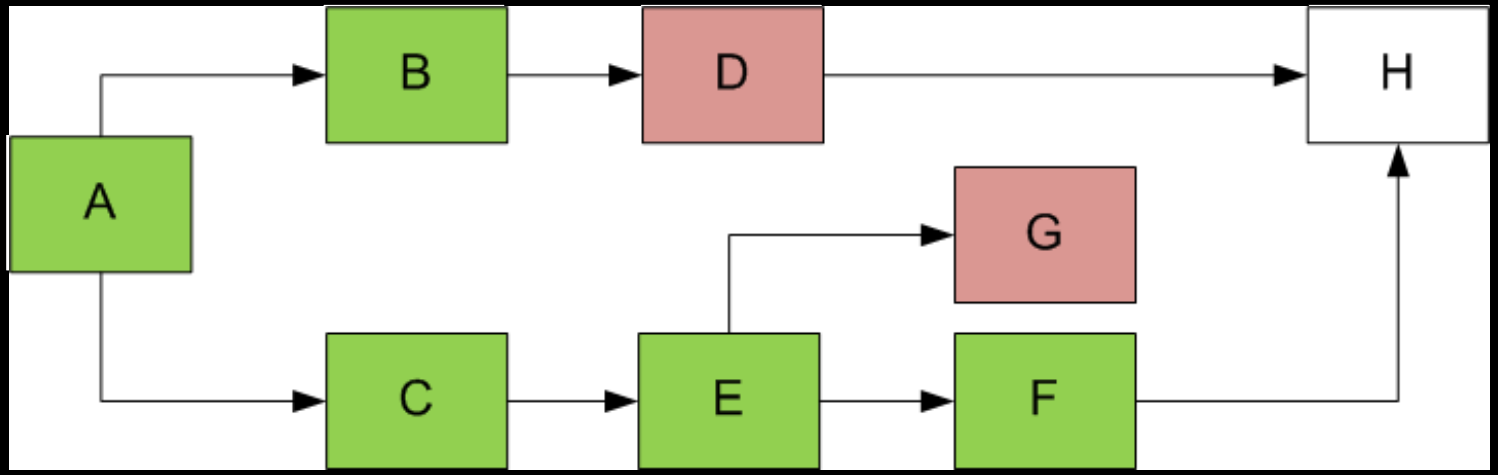
Background: Multipath

- Measurement Pattern



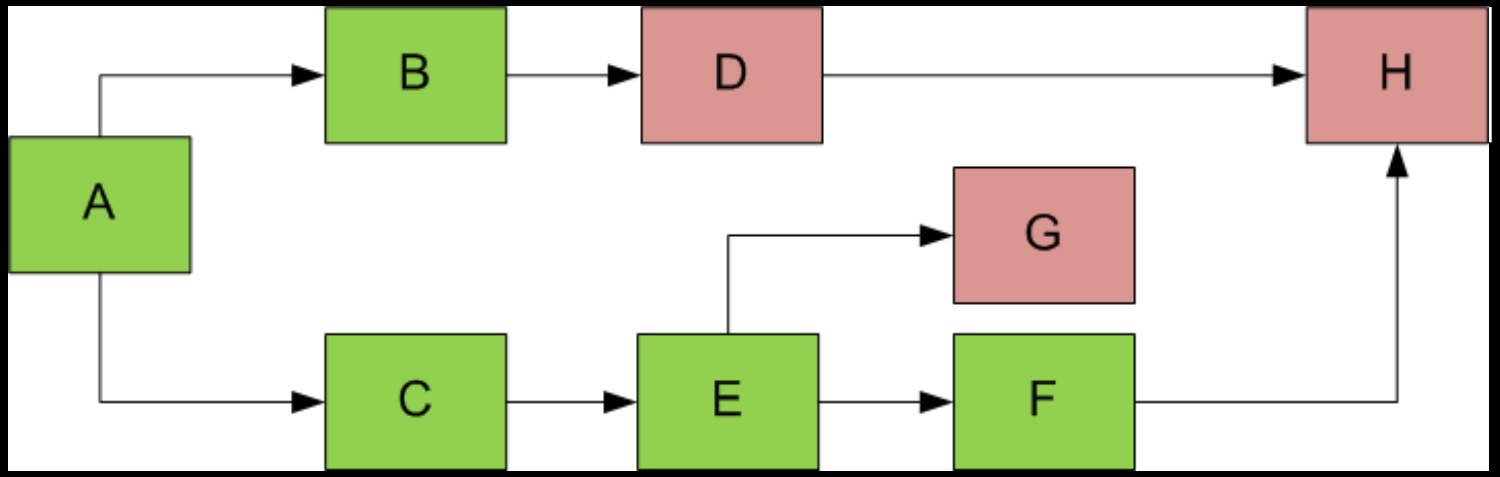
Background: Multipath

- Measurement Pattern



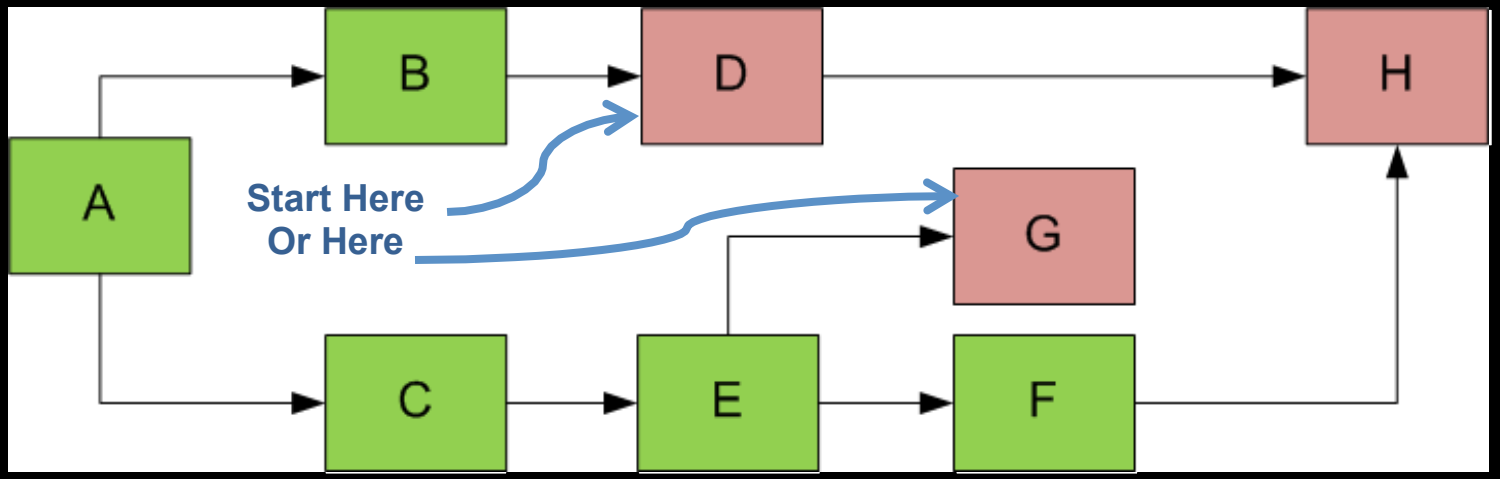
Background: Multipath

- Measurement Pattern



Background: Multipath

- Instructional Recommendation



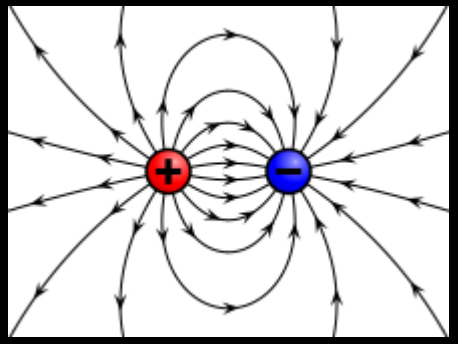
Item and Test Analysis Data for the Study

- 10,000+ unique science items
- 150+ unique test forms
  - All forms had high internal consistency (Coefficient Alpha of 0.80 or greater—used Spearman-Brown correction for tests with fewer than 20 items)
- 50,000,000+ test taker responses

Assessment Item Alignment Data

- 8,400 fine grained indicators (topic or skill labels)
- 331 closely related groups of indicators into teaching units (subtopics)

Example Alignment



- Subtopic: Electric Field Lines
  - Learning Indicators:
  - *Recognize* that lines of force are imaginary lines that represent how a positive test charge would move in the presence of the source charge, field lines also indicate the relative strength of the electric field at a given point in the space of the field
  - *Recognize* that negative charges have electric field vectors that radiate inward toward (that is point to) the charge
  - *Recognize* that positive charges have electric field vectors that radiate outward (that is point away) from the charge
  - *Recognize* that when the force between two charges is attractive, their respective vectors must have opposite signs
  - *Recognize* that forces exerted by electric fields are repulsive if the stationary test charges and stationary source charges are like
  - *Recognize* that when the force between two charges is repulsive, their respective vectors must have the same sign

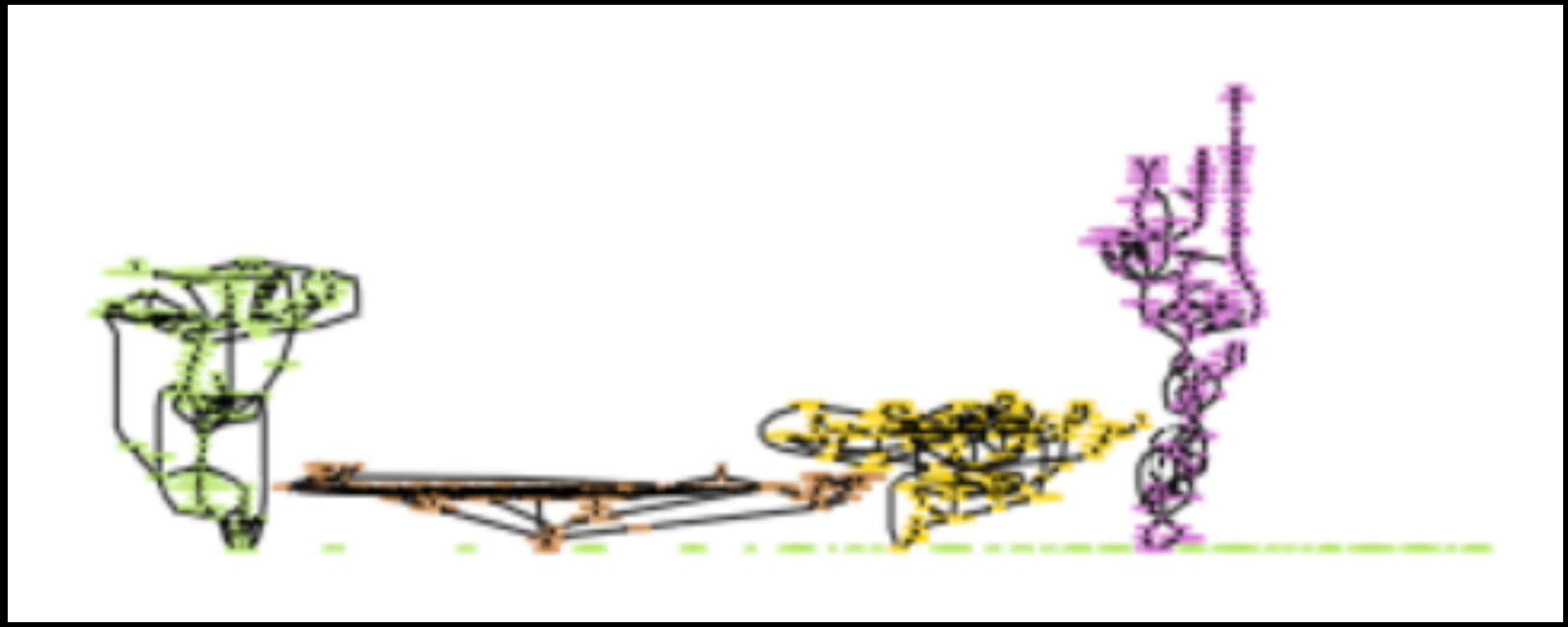
Image from: Electric field. (2011, October 11). In *Wikipedia, The Free Encyclopedia*. Retrieved 20:06, October 12, 2011, from //en.wikipedia.org/w/index.php?title=Electric_field&oldid=455060612

Method

- Phases
  - Trim alignments based on current paths
  - Compare median difficulty of aligned items
  - Adjust
    - Links
    - Item alignment
    - Subtopic definitions
    - Learning indicators alignment

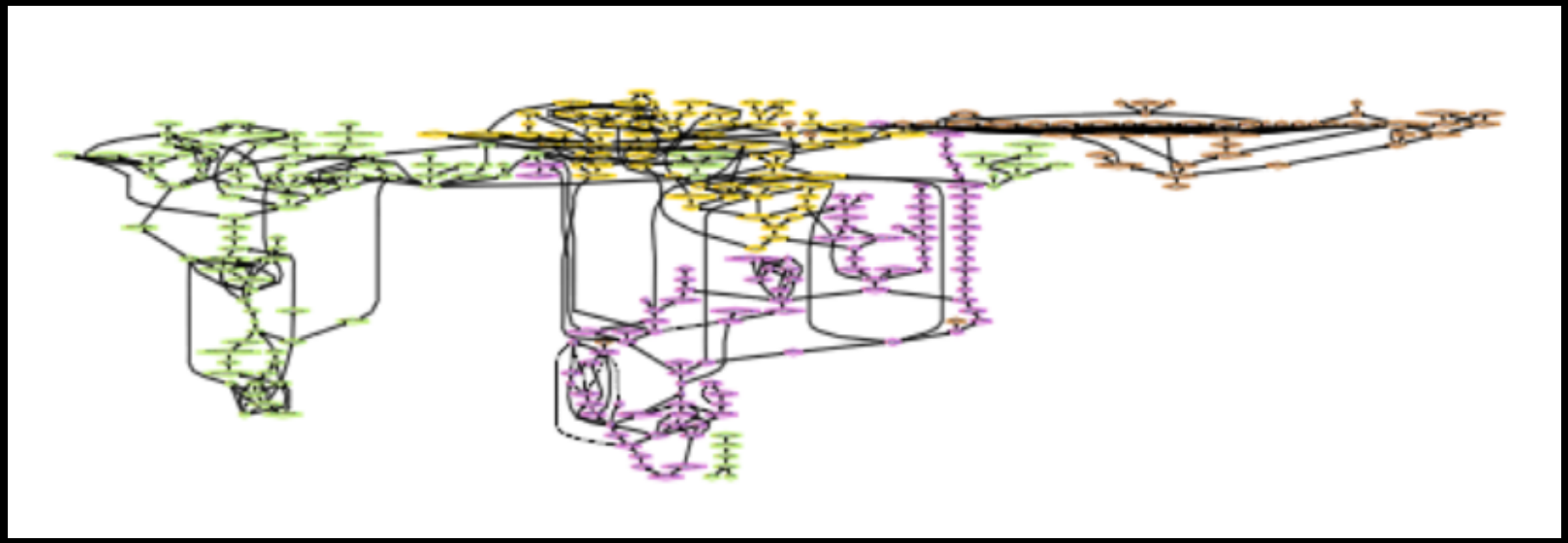
Phase Results

- In the beginning, the paths were bound within subjects



Phase Results

- phase 2…connected the subjects

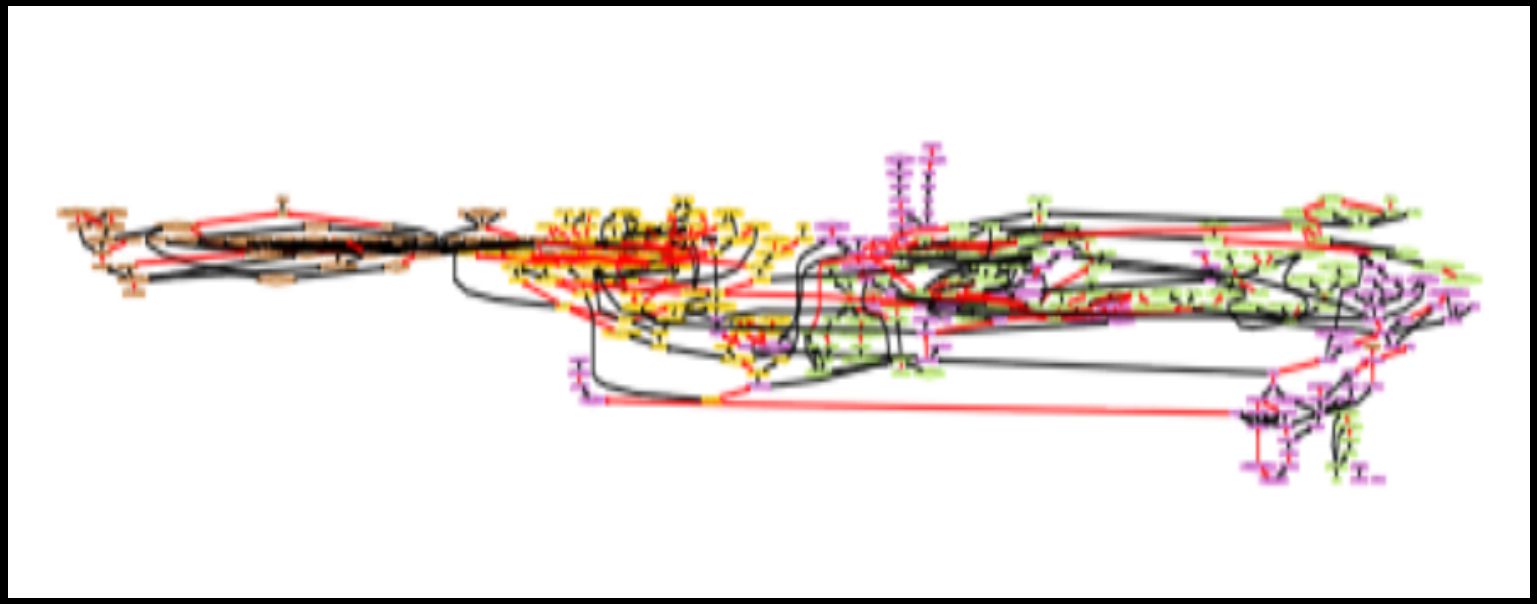


Phase Results

- …phase 3… added the median p-values to the subtopics on the graph

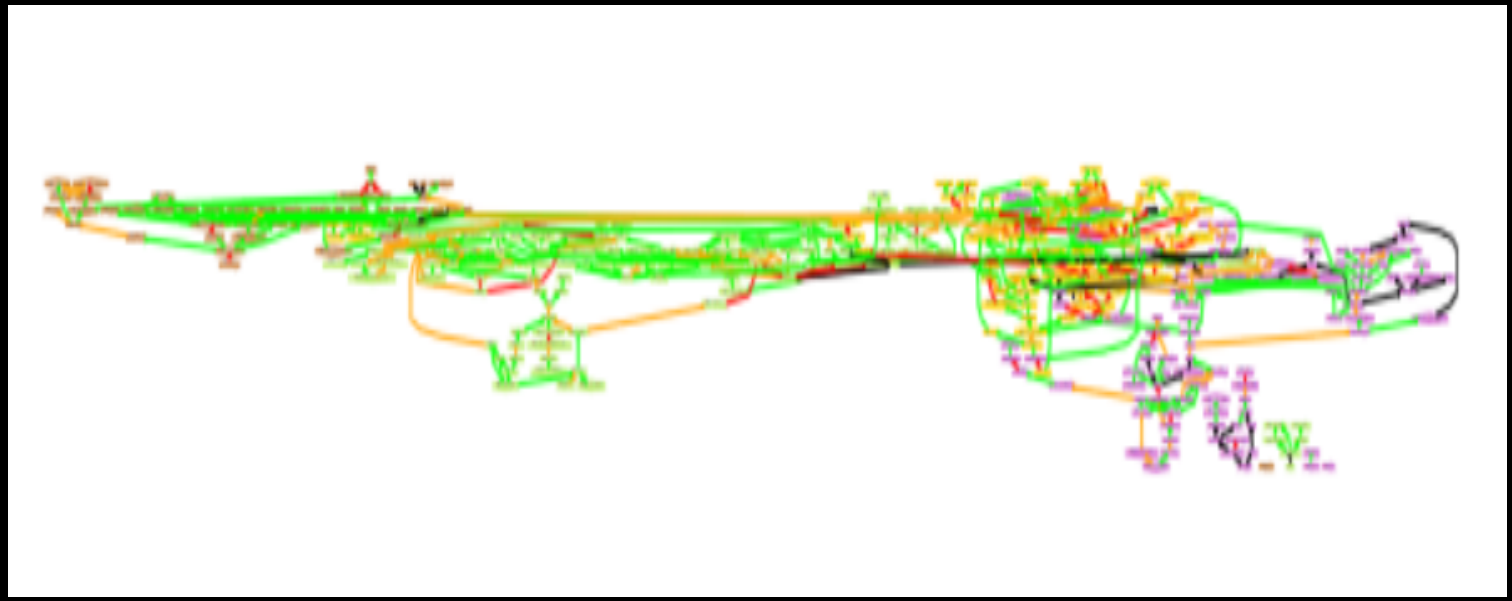
Phase Results

- phase 5…added color coding on the relationships



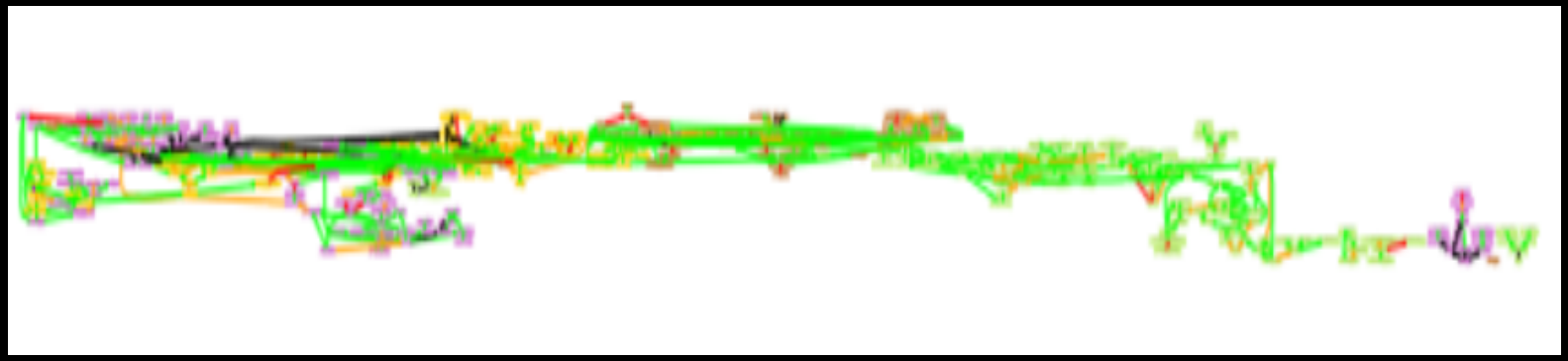
Phase Results

- phase 6…realigned some items…more green… headed the right way!



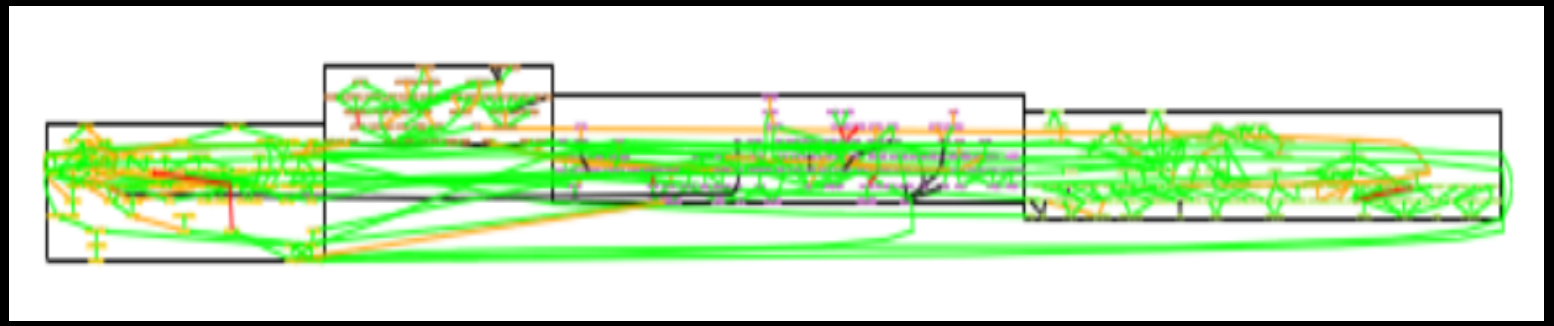
Phase Results

- phase 9…added the percentage of valid links in each direction from the subtopic



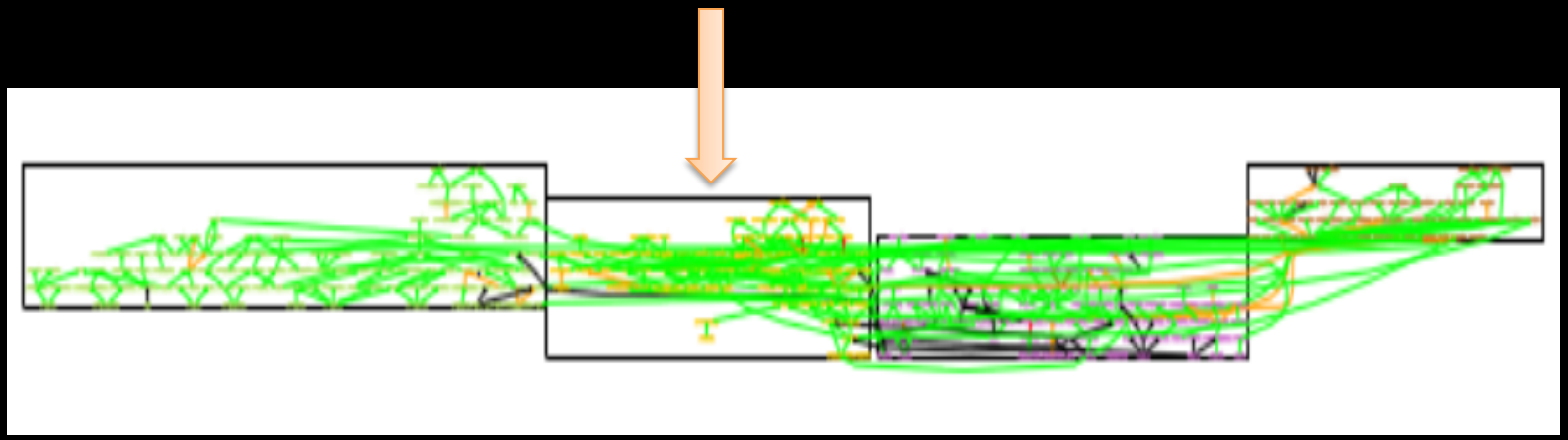
Phase Results

- phase 19…grouping subjects together



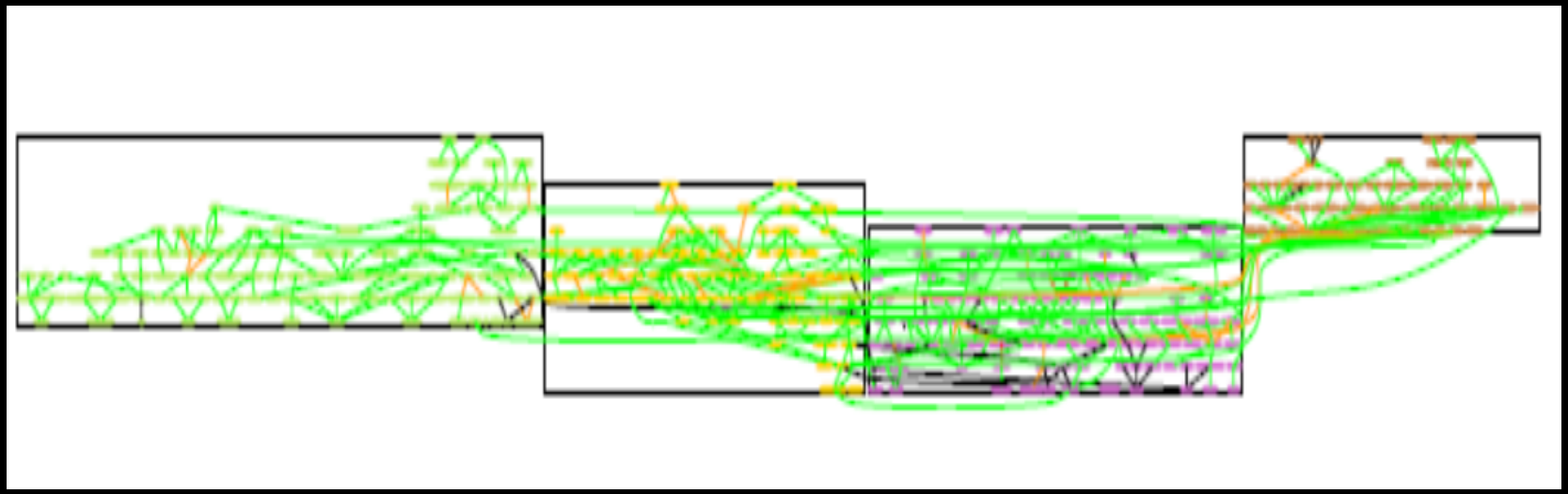
Phase Results

- phase 23…reorganized concepts grouping in one of the subject areas…some issues resolved.



Phase Results

- Phase 25… a few additional tweaks and done!

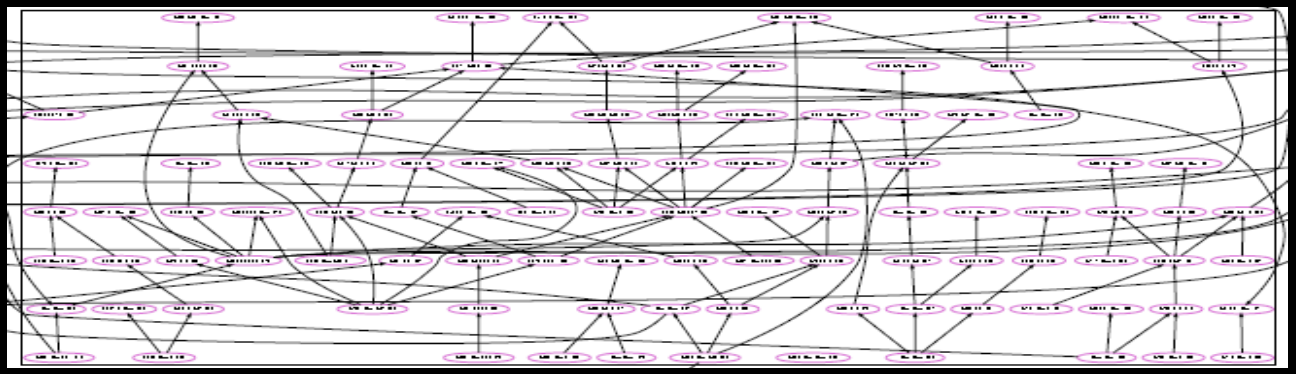


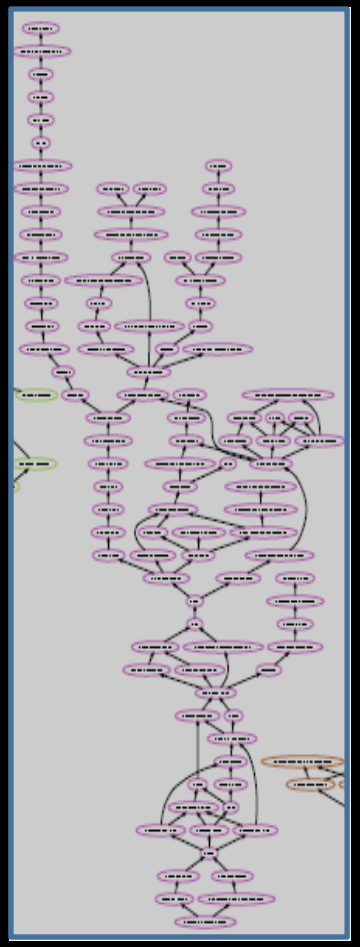
Findings

- Visualizations helped to identify how to update the paths
- Changes in both links and alignment must be considered to reconcile data with SME intuition

Findings

- The paths get shorter, and more complex





Using routinely-collected big data to predict student dropouts

- Question: How do student demographic and academic characteristics relate to dropout rates?
- This study is part of a larger plan to create "research pipelines"

Research Context

- Study conducted in an online university that offers ~1000 different online courses
- University has an extensive database of information on students and their performance, including measures of learning, student satisfaction, postgraduate success, etc.

Analyses

- Analyzed academic and demographic characteristics of degree-seeking students (*N* = 14791) enrolled during a two-year
- Survival analyses (a type of logistic regression) revealed that measures of student performance were significant predictors of the likelihood of dropping out.

Dropout results, 1

Predictor	Effect on dropout risk
Age between 29 and 38	Reduces risk by 40%
Age between 38 and 45	Reduces risk by 7.5%
Older than 45	Increases risk by 30.7 %
Has transfer credits	Increases risk by 236 %
Enrolled in 200 level courses	Increases risk by 7.4%
Enrolled in 300 level courses	Reduces risk by 26.8%
Enrolled in 400 level courses	Reduces risk by 66%

Dropout results, 2

Predictor	Effect on dropout risk
In the military	Reduces risk by 72.8%
Previous college education	Increases risk by 13.3 %
Female	Increases risk by 330 %
Estimated financial contribution from family	Increases risk by 17 %
Married	Reduces risk by 35.7%
Has dependents	Reduces risk by 58%

Four (Adaptive) Solutions for Motivation Problems

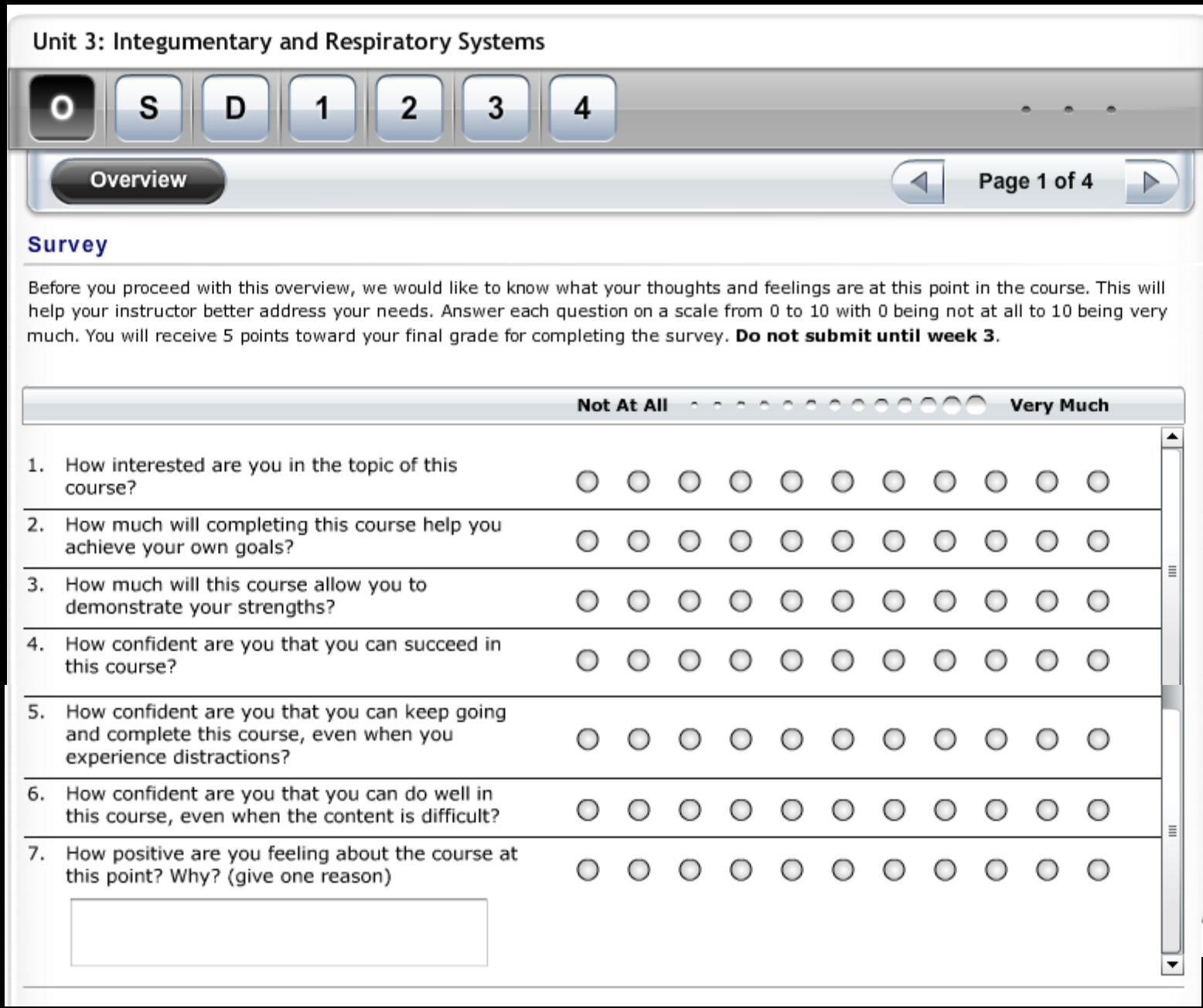
1. Increase the personal value of a learning task
   - If value is low, no starting or persisting
2. Increase (or decrease) student's confidence or self efficacy
   - Very low or high confidence decreases persisting and effort
3. Provide more positive emotional climate (mood)
   - Negative emotions wreck starting, persisting and effort
4. Reframe students beliefs about the controllability of problems
   - When students believe that problems are uncontrollable they will quit or manufacture an excuse not to start or persist

Survey of beliefs and emotion (value, confidence, mood)

Value
- Intrinsic
- Utility
- Strengths

Confidence
- Success
- Distractions
- Difficulty

Mood
- Positive feeling



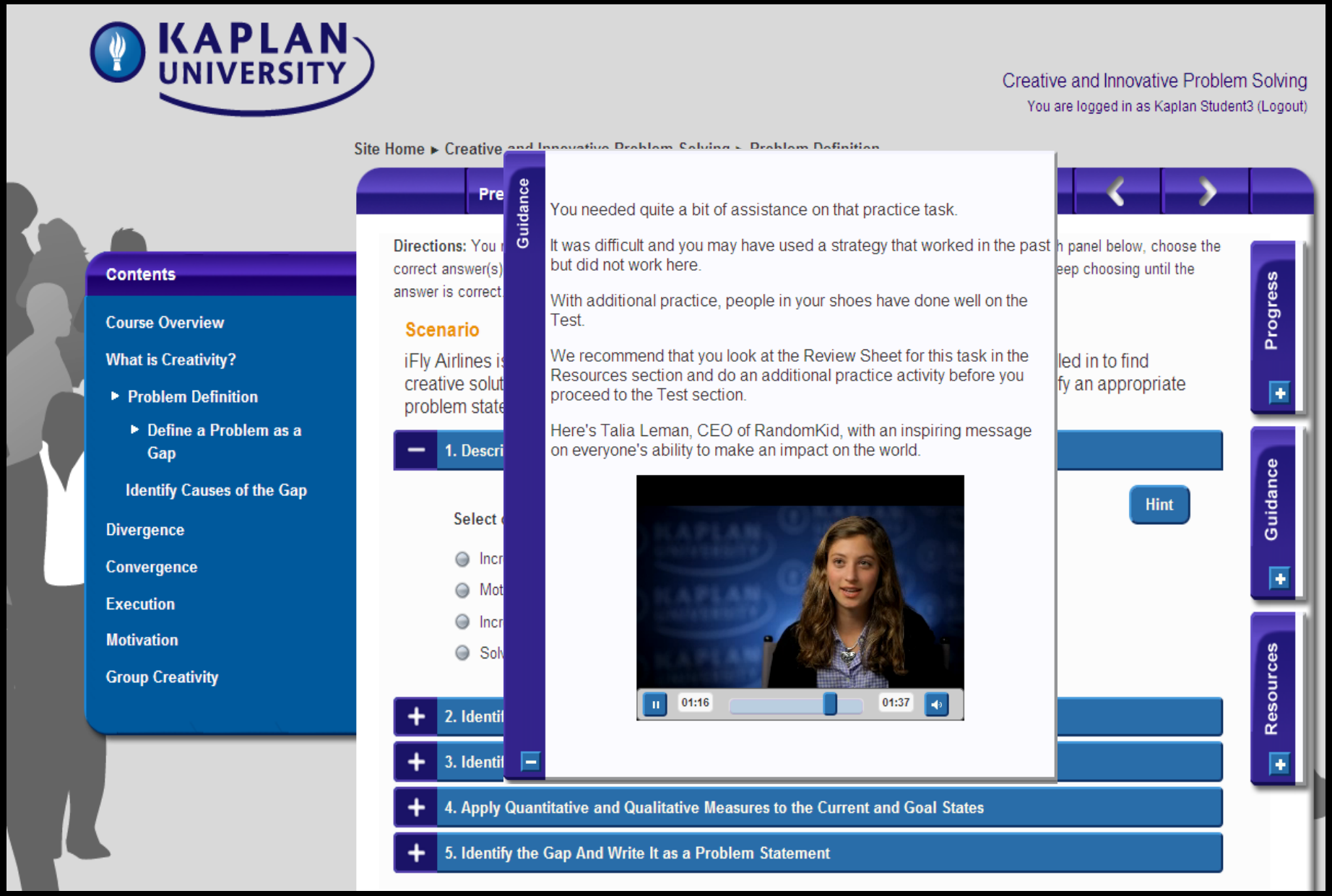
Rules and guidance for different patterns of performance and motivation

Perceived Value	Guidance for students
Low	Explain and model the value (benefits, risks)

Performance	Self-Efficacy	Attribution	Guidance for individual students
High	All	NA	Proceed to Test
Low	High	Controllable	Reduce confidence, increase effort
Low	High	Uncontrollable	Reattribute to effort or different strategy
Low	Low	Controllable	Boost confidence
Low	Low	Uncontrollable	Reattribute to effort or strategy and boost confidence
Low	Med	Controllable	Motivation OK. Recommend more practice
Low	Med	Uncontrollable	Reattribute to effort or different strategy

Performance	Outcome Session Time	Guidance for individual students
Low	Low	Focus on spending more time
Low	High	Focus on using time more productively – provide training in study skills

Guidance: Original concept course with Remote-Learner (2010)
Dynamic guidance based on patterns of performance and motivation



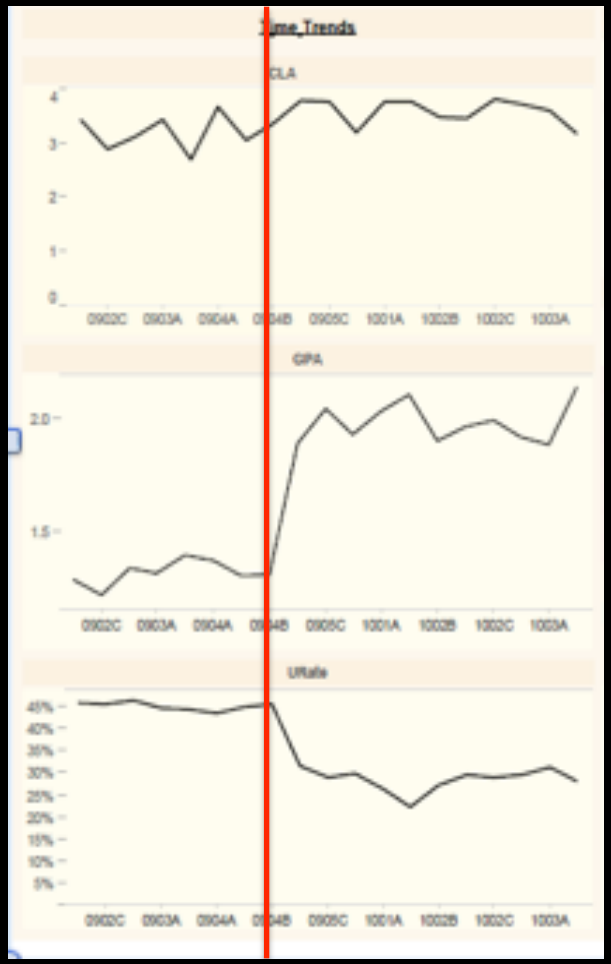
Technology affordances for learning analytics

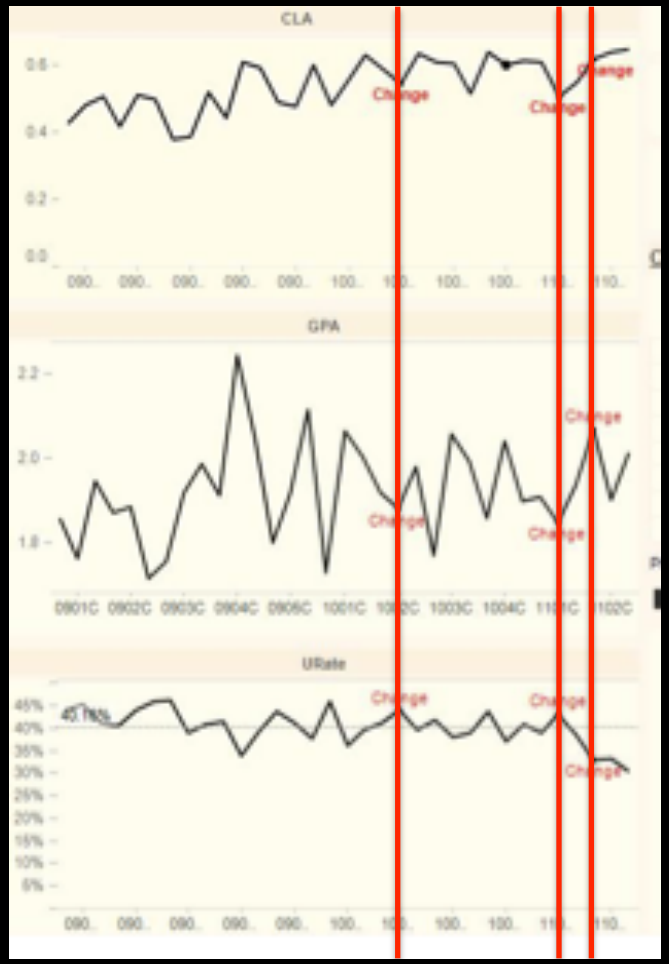
- Longitudinal analyses of big data, including quasi-experimental studies
- Micro data
- Experimental investigations
- Adaptive instruction, personalized learning

"After we got the computer, we were better able to track our problems. At least, that's one interpretation."

Longitudinal analyses

Monitoring trends, perturbations





As valuable as these analyses of routinely-collected big datasets may be…

They don’t tell us how to help students who aren’t succeeding

We need to:

- Better understand *why* students struggle (through in-depth qualitative studies)
- Determine which combinations of indicators and performance patterns constitute “red flags” requiring immediate, strong intervention and which may call for less intensive strategies
- Test new instructional, motivational and support strategies to help students who are having difficulty and likely to drop out

Building research pipelines

Running the pipelines

- Some large classes have 50-100 sections with 20-30 students in each section
  - Students can be randomly assigned to sections
  - Sections can be randomly assigned to treatments
- In some cases, students within sections can be randomly assigned to treatments

Who's using the pipelines? Partnerships with external researchers

Studies currently underway or planned with:

- John Sweller (worked examples)
- Dick Clark (motivational priming, Cognitive Task Analysis)
- Rich Mayer (multimedia design)
- Jeroen van Merrienboer and Marriette van Loon (self-evaluation of knowledge)
- Jan Cannon-Bowers and Valerie Shute (games to improve persistence)
- Todd Rogers (motivational priming)
- Khan Academy, EdX

Question

How will results from longitudinal big data analyses compare to experimental results?

How we use Mechanical Turk

A global pool of workers available to do tasks based on instructions.

Collect a dataset that doesn't now exist	Log into a system and use an asset	Attempt to perform a particular task on one of more websites	Take a test and be paid per correct answer
Take a test for use in equating, or identifying bad items	Provide guerrilla feedback on an interface mockup	Identify cultural differences in perception of an interface	Generate content

Global pool. This map shows a 10% sample of workers (red) available on Amazon.com's Mechanical Turk.

HUMAN SUBJECT RESEARCH

Social Science for Pennies

Social scientists are turning to online retail giant Amazon.com to cheaply recruit people around the world for research studies

It's a problem that all social scientists face. You have a brilliant idea for a study. You have the experimental design all worked out, and your university's review board has approved it. But you still have to recruit hundreds of people as subjects for the experiment.

Gabriel Lenz, a political scientist at the University of California, Berkeley, faced this problem last year when he and collaborators wanted to follow up on another group's study of voting behavior (*Science*, 10 June 2005, p. 1623). For that study, Americans were shown photographs of past U.S. congressional candidates and asked to rate the politicians on various characteristics, such as competence and attractiveness. Even though the study subjects had no information beyond an image of the candidates' faces, their snap judgments were a significant predictor of who actually won the races. Lenz wanted to see if that surprising result collapsed when those evaluating the photos come from cultures different from those of the candidates. But how to recruit people currently registered with the MTurk site as available for work. The task of rating the political candidate photos required about 4 minutes. "We played around with various payment rates," Lenz says. For Turkers based in India, the researchers started low, offering 15 cents. In just 4 days, they received data from 100 people. Then for a control group, they recruited more than 300 Americans for between 20 and 50 cents each. The total cost? About $160, and that includes the 10% fee Amazon charges.

In just a few weeks, Lenz had all the data his group needed. In spite of the cultural differences, the snap-judgment effect persisted: American and Indian subjects predicted the winners of Brazilian political races based on nothing more than a mug shot, the researchers reported last year in the social science journal *World Politics*.

As others follow Lenz's lead, many more social science papers using MTurk will appear in the coming years, predicts Adam Berinsky, studies, Turkers get paid only if they generate usable data. This is necessary to eliminate not only people who don't understand the task but also "spammers," people who try to exploit MTurk by skimming through the jobs and giving random responses wherever possible to accelerate the process.

For example, Lenz had to reject about 20% of his American and 50% of his Indian Turkers for those reasons. But that is a manageable problem, Berinsky says. A counterintuitive solution is to keep the price low. "If you offer more than a dollar, you attract the spammers who sort jobs by level of pay," he says. "You have to find the sweet spot where the payment is not too high but still attractive enough for most Turkers." So far, that sweet spot seems to be between 15 and 50 cents for a 10-minute job.

Even if MTurk is cheap and fast, doubts will linger about interpreting data from research subjects whom you never meet. To address those concerns, Berinsky and Lenz are teaming up with Gregory Huber, a political scientist at Yale University, to study the Turker population. And of course, they are using MTurk to do so. They recently replicated two classic survey experiments and a political science experiment. In each case, the data obtained with MTurk were consistent with published studies that tested people in laboratories.

The scientists have found some differences, too. Turkers "are younger and more ideologically liberal than the U.S. public," Berinsky says. However, they are more repre-

What if you could conduct an experiment and get results every day?

One-day Mechanical Turk study

Condition	Average Posttest (Out of 12)	Standard Deviation	Median Instructional Time (Minutes)	N
Study 8 worked examples	6.36	2.97	8.15	153
Study 15 worked examples	5.84	2.97	12.87	148
Video Instruction	4.73	2.79	99.32	107
Test Only – No instruction	5.07	2.67	NA	84

A oneway ANOVA showed a significant effect for different instructional conditions on posttest performance, $F(3,488) = 8.08$, $p < .0001$, $\eta 2 = .05$.

Looking forward: technology options

- Simulations
  - Good for complex situated problem solving
  - Enable clickstream analysis
  - Key issue: high v. low fidelity
    - High fidelity = closer to real-life context (e. g., job), but more expensive to develop
    - High fidelity complex situations not ideal for novice learners – too overwhelming
- MUVE - multi-user virtual environments
  - Can track every action and thus evaluate individuals, which can be difficult if you just look at a team outcome
  - Intrinsic feedback (responses to actions) or after action reviews
  - Good for assessing teamwork skills
- Social media
- Assessments embedded in online tutors or other adaptive programs
- Embedded measures of motivation, beliefs moods
- Adaptive assessments – ultimately drive instruction
  - Bayes-net-driven: Catalyst
  - IRT-driven: KHEC, GMAT

Contact

dniemi@kaplan.edu