

A Log Is Not Evidence: Designing Interactive Tasks for Process-Based Assessment

Engin Bumbacher



University of Teacher Education Lausanne, Switzerland



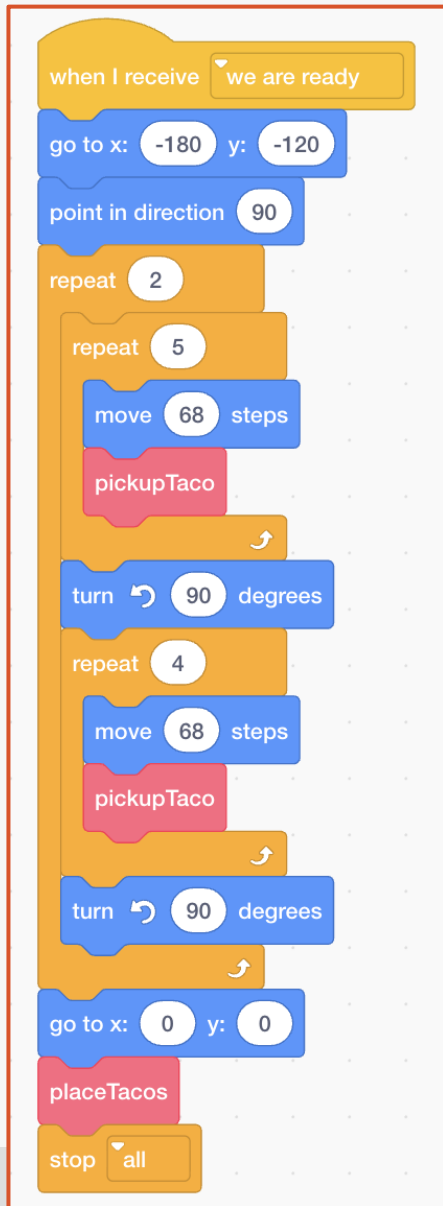
OECD PISA Expert Group PISA Learning in the Digital World 2025

From process traces to evidence: theorizing the task–evidence mechanism

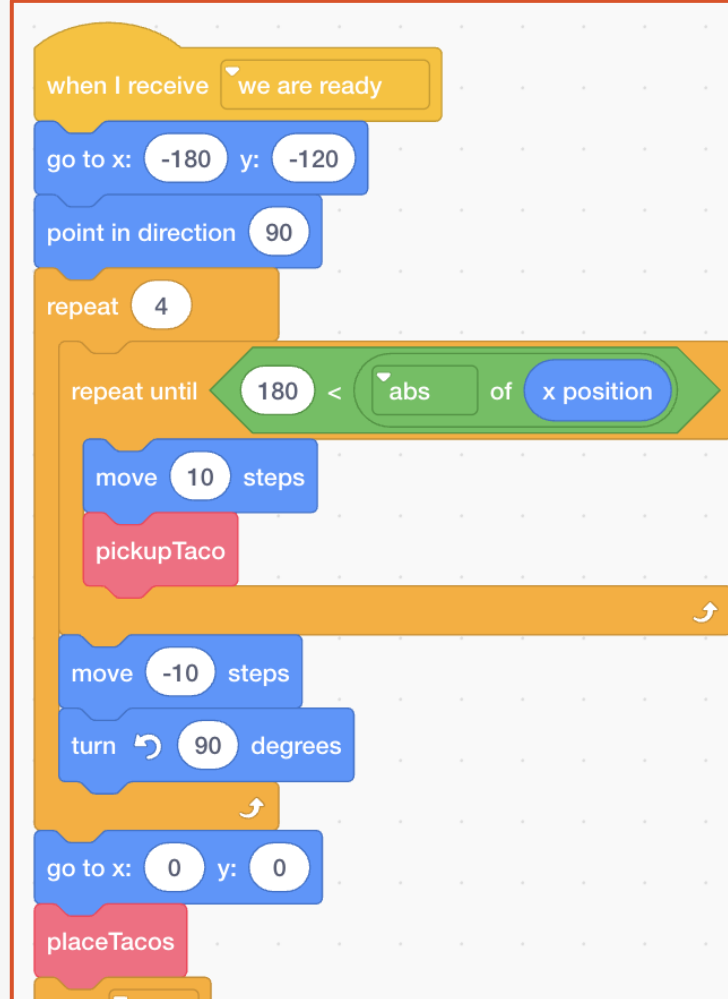
I'll bring together several strands of my work to examine how task and tool design shape the processes we observe, and what this means for evidence construction and interpretation.

- What have we learned from empirical studies?
- How can we theorize the task more richly?
- What does this mean for future work?

Student A Final Program

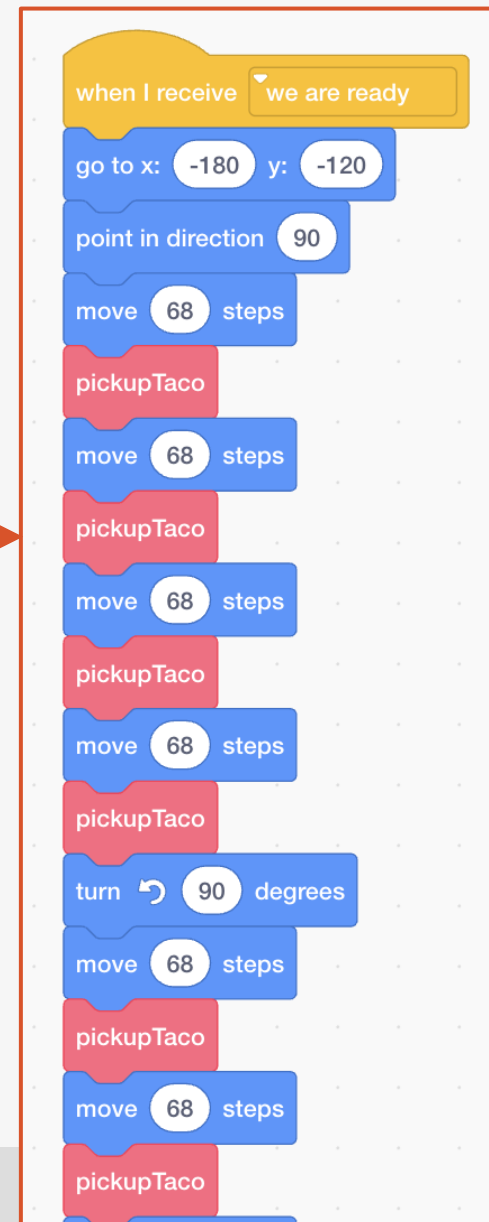


Student B – Second-to-Final Program



... make otherwise invisible aspects of competence visible.

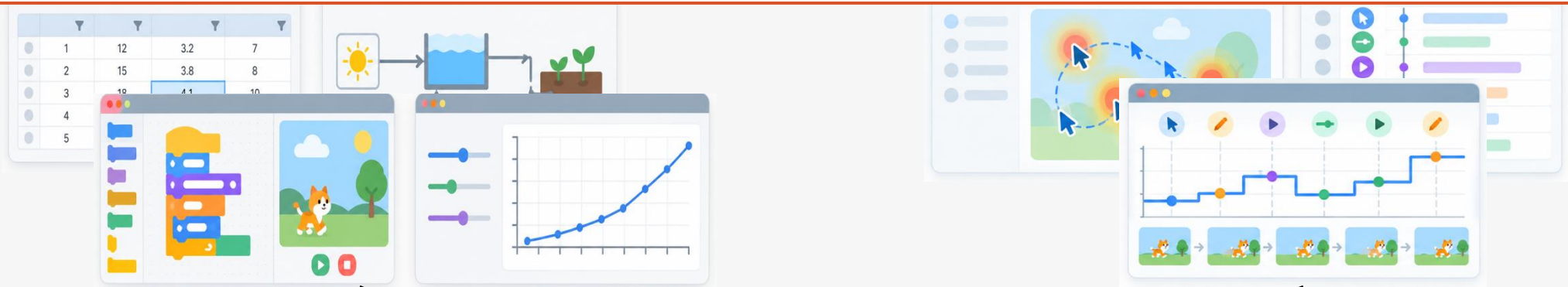
Student B – Final Program



From process data to evidence

Interactive Tasks

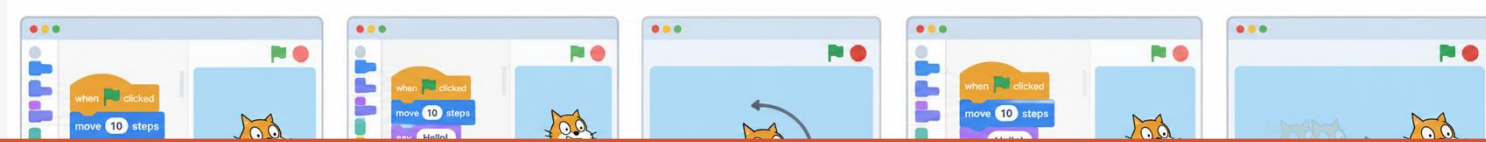
Create decision opportunities and express competence through decision-making
Wieman & Price (2023)



New opportunities... and new challenges

Evidence about learner's competence

From process data to evidence



Does this sequence provide evidence of «systematic testing»?



Run



Modify



Run



Modify



Run

Task Variant A:

«After every code edit, run your program and inspect the results.»

Task Variant B:

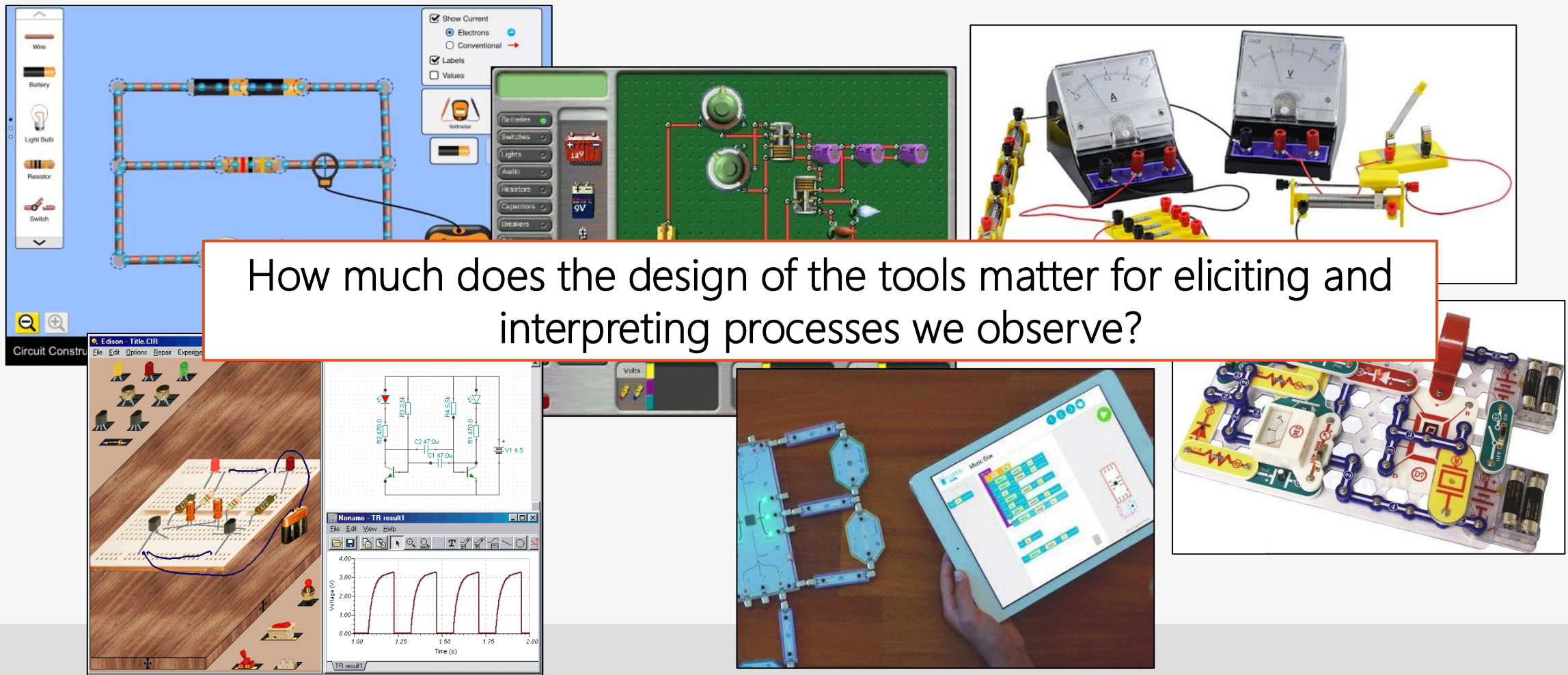
Testing is available, but not explicitly prompted.

Interpretation problem

The task design shapes what an observed process means. How?

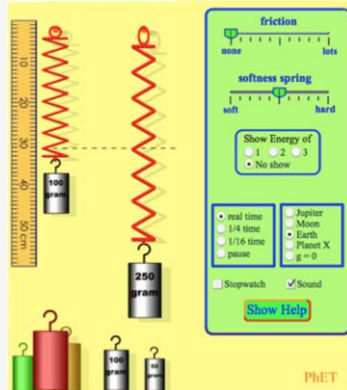
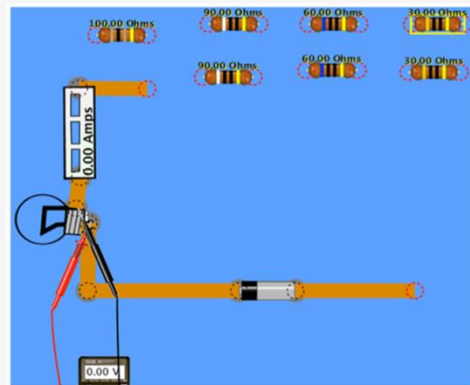
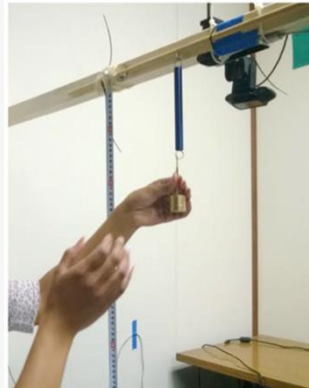
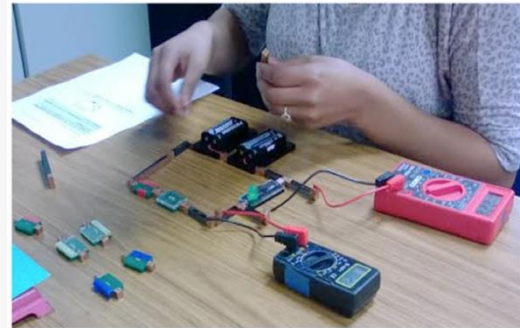
And interactive tasks have to be enacted in an interactive tools...

Large variety of tools for same domains...



How task design changes how learners engage in the interactive task (in science inquiry)

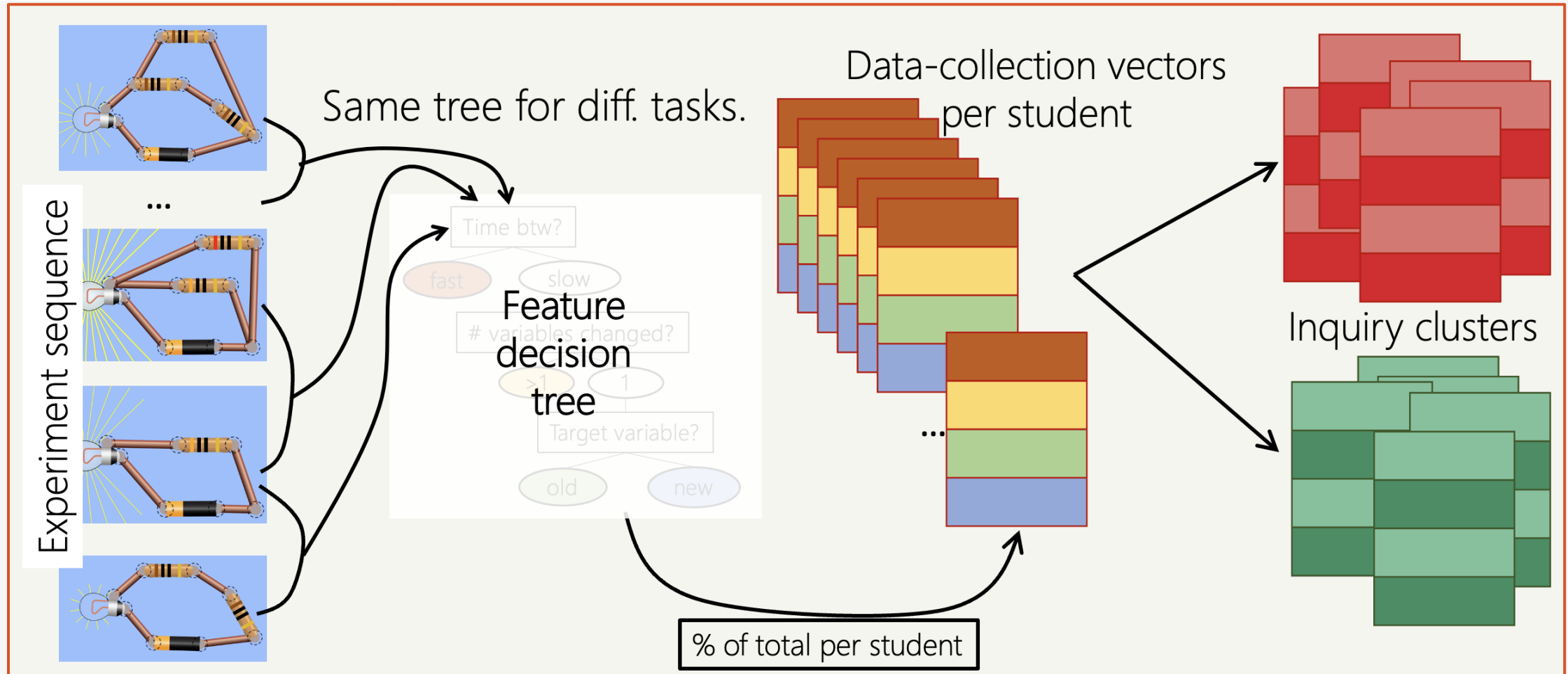
Bumbacher et al. (2018)

OELE type		1. Mass & Spring		2. Electric Circuits			
Same higher-level inquiry problem	Virtual	Pre-test		Post-test	Pre-test		Post-test
	Physical				Pre-test		

70 US community college students

How task design changes how learners engage in the interactive task (in science inquiry)

Bumbacher et al. (2018)

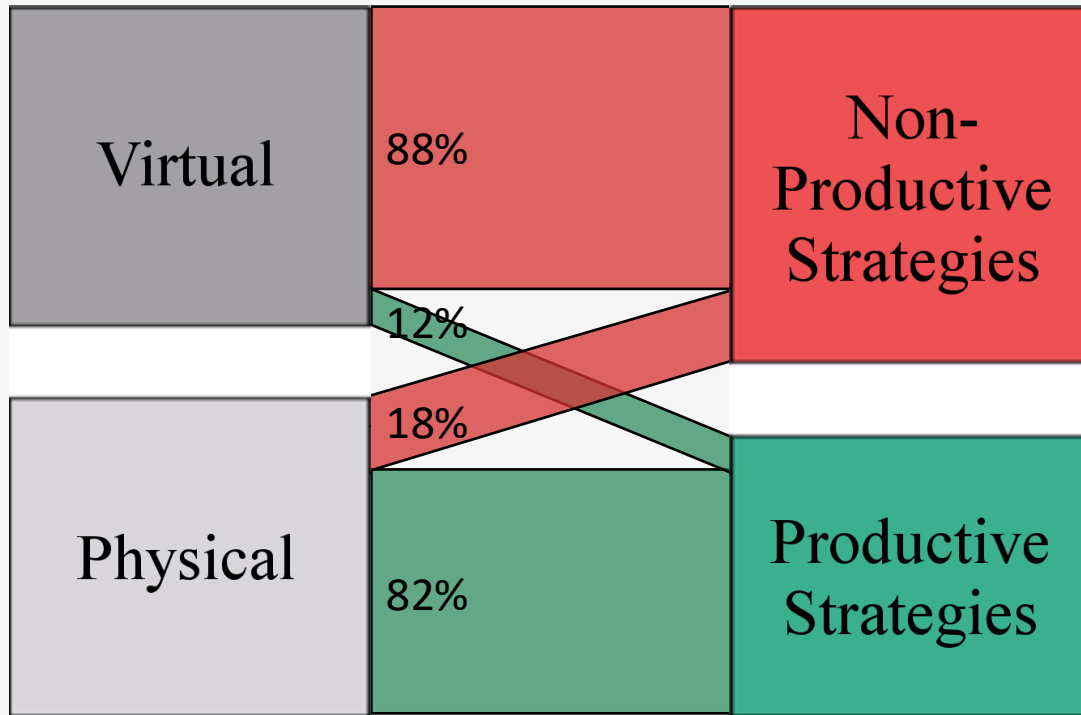


70 US community college students

How task design changes how learners engage in the interactive task (in science inquiry)

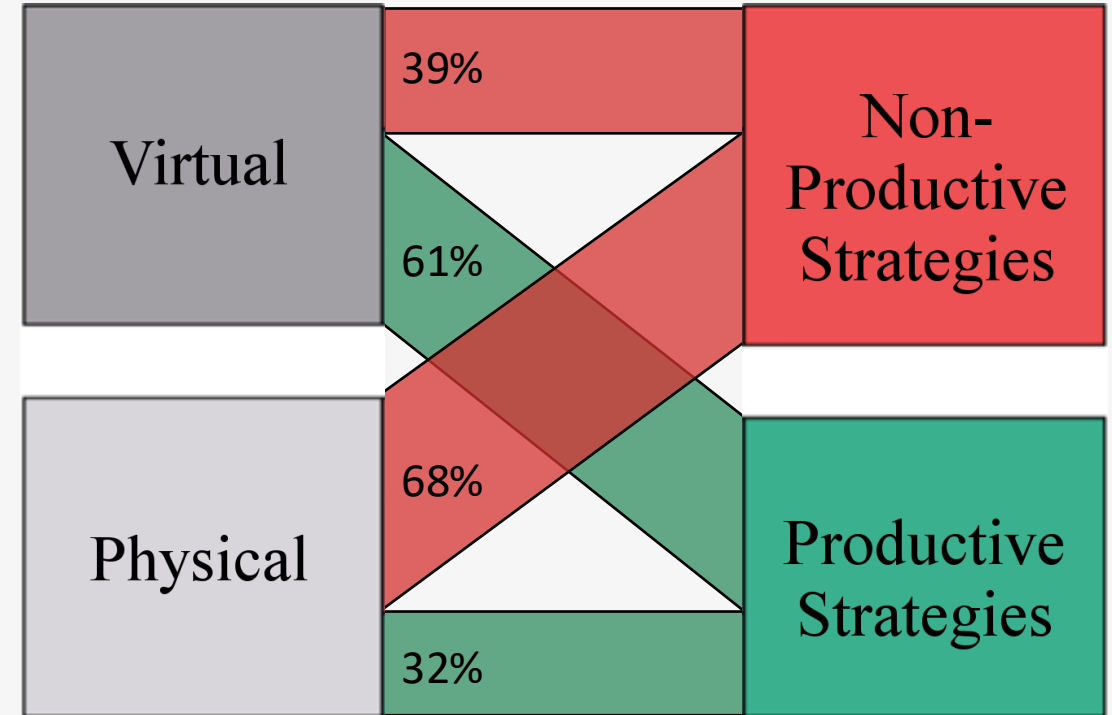
Bumbacher et al. (2018)

Mass & Spring



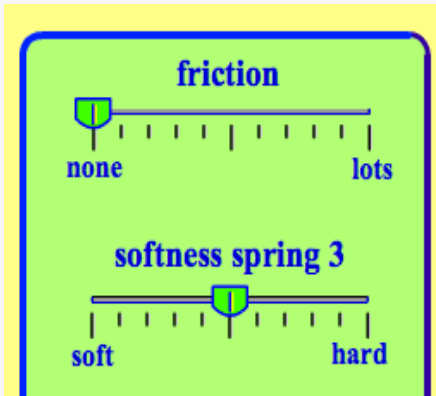
$p < 0.001$

Electric Circuits

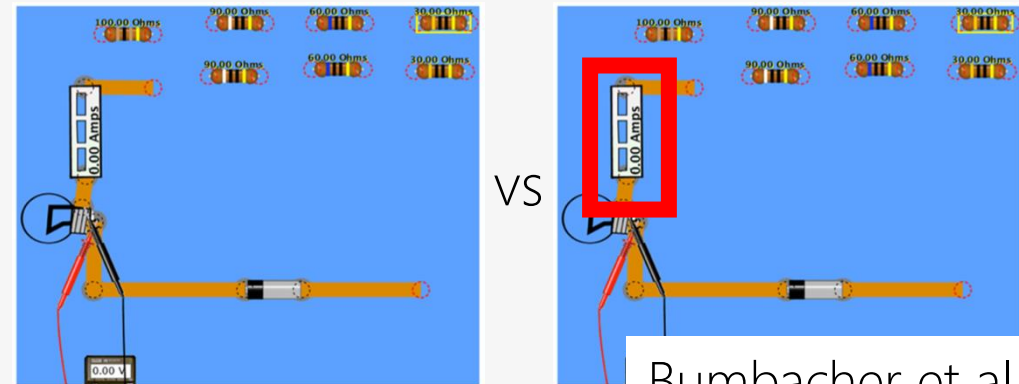


$p = 0.02$

Tool design shapes at the feature level what becomes observable and what the observable can mean



Change **one design affordance** → different elicited process

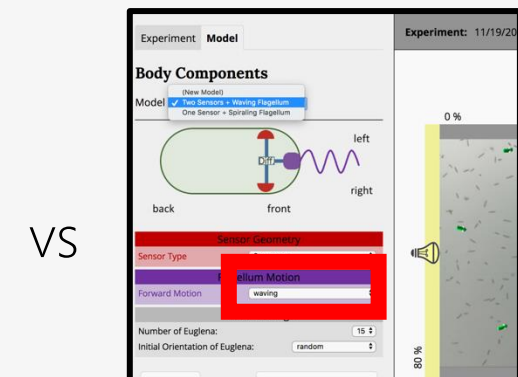
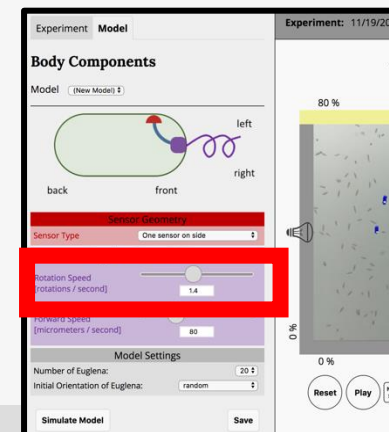


Same **decision opportunity**, different **interactive realization**.

rapid vs slow manipulations of same variables

→ different **elicited process**

→ different interpretation



Bumbacher et al. (2019, 2020)

Task and tool design shapes at the feature level what becomes observable and what the observable can mean

The Elicitation Problem

The task and tool design shapes which processes learners actually generate and make observable.

Same learner, same higher-level task, different design feature

- Different interaction realization
- Potentially different process

The Interpretation Problem

The task and tool design shapes what an observed process means.

Same process, same higher-level task, different design feature

- potentially different evidence

Accounting for task design when using process data to infer competencies

REVIEW ARTICLE

British Journal of Educational Technology BERA

A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices

Karen D. Wang¹ | Jade Mai Cock² | Tanja Käser² | Engin Bumbacher³

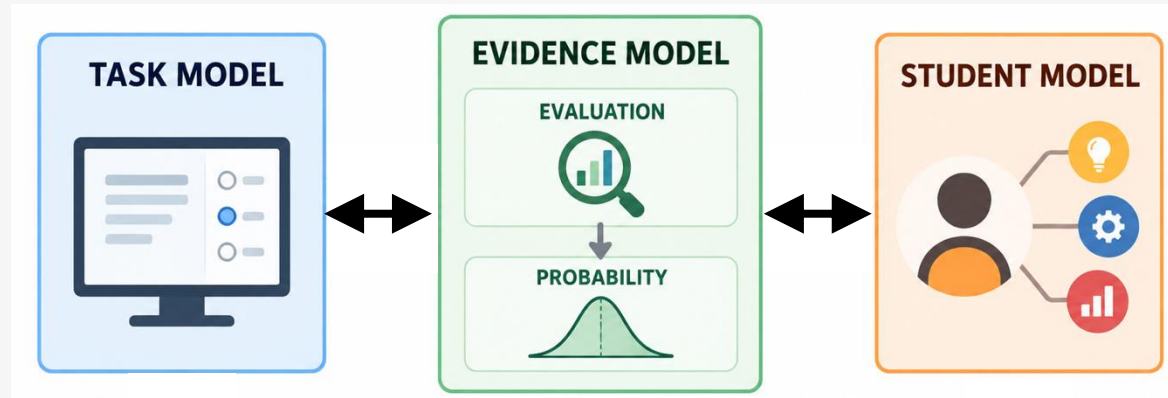
- 70 studies of log data in open-ended learning environments (OELE) for STEM learning



- Little systematic incorporation of task-design differences into construction of process evidence
- The Generalization Problem:
How should task and tool design enter the inference from task-specific behavior to general competency claims?

ECD gives the architecture: How to represent task features to account for their consequences?

Evidence-Centered Design (Mislevy et al., 2005)



The task model specifies features and conditions of situations designed to provide decision opportunities and elicit the targeted evidence.

- **Example:** Same CS competency (e.g., applying an if-then structure) ECD-based assessment design can elicit evidence through selecting an appropriate structure or constructing code in an open-ended problem. The task form determines how the competence can become observable. (Newton et al., 2021)

ECD gives the architecture: How to represent task features to account for their consequences?

How richly must we theorize the task for process evidence?

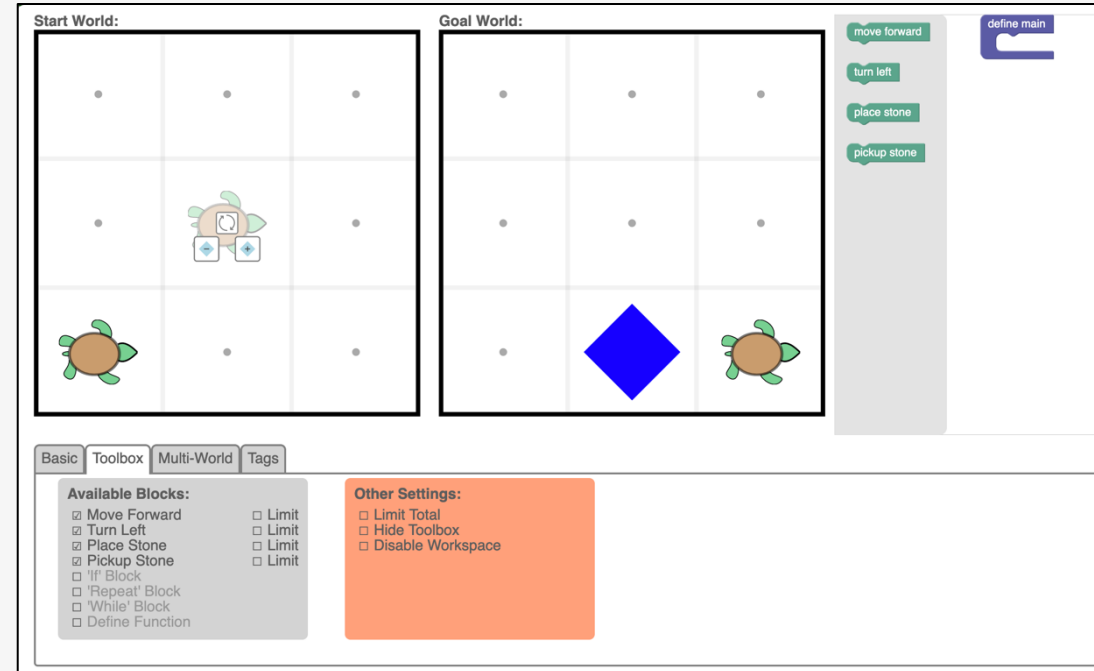
Task design

- what **decision opportunities** does it provide?
- how are those opportunities **interactionally realized**?
(slider, dropdown, physical manipulation, available controls, action sequence,...)
- what process does that configuration **elicit**?
- how can that process be **interpreted** as evidence?

Developing that account requires both theory and empirical work.

Configurable tools to empirically probe within-tool the consequence of task design on process

Bumbacher et al. (2024; in review)



The Platform for Innovative Learning Assessments (PILA)

Research-empowered tools for innovating assessments of 21st century competences



[Learn More About PILA](#)

<https://pilaproject.org/>

Computational Thinking (CT) : From an established pen-and-paper test to interactive problem solving

Bumbacher et al. (2024; in review)

Which instructions take 'Pac-Man' to the ghost by the path marked out?



Option A

```
repeat 4 times
do
  repeat 3 times
  do
    move forward
  turn right 90°
move forward
```

Option B

```
repeat 3 times
do
  repeat 4 times
  do
    move forward
  turn right 90°
move forward
```

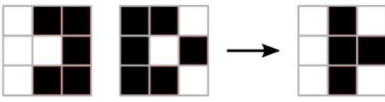
Option C

```
repeat 3 times
do
  repeat 4 times
  do
    move forward
  turn right 90°
move forward
```

Option D

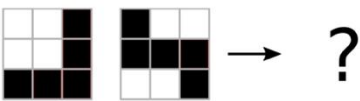
```
repeat 4 times
do
  repeat 3 times
  do
    move forward
  turn right 90°
move forward
```

Combining Card A and Card B, you get Card C:



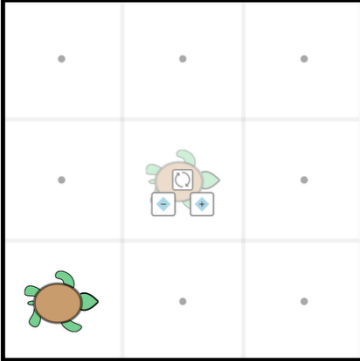
Card A Card B Card C

Question:
How many black cells will Card F have after combining Card D and Card E?

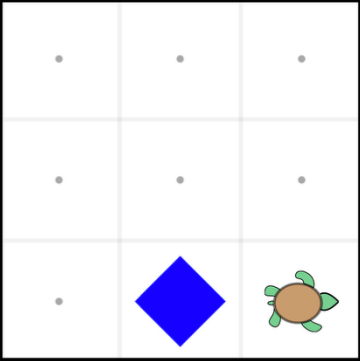


Card D Card E ?

Start World:



Goal World:



move forward
turn left
place stone
pickup stone

defining main

Basic Toolbox Multi-World Tags

Available Blocks:

- ☒ Move Forward
- ☒ Turn Left
- ☒ Place Stone
- ☒ Pickup Stone
- ☐ 'If' Block
- ☐ 'Repeat' Block
- ☐ 'While' Block
- ☐ Define Function

Other Settings:

- ☐ Limit Total
- ☐ Limit
- ☐ Hide Toolbox
- ☐ Limit
- ☐ Disable Workspace

B+CTt: 25-item paper-based assessment (Wiebe et al., 2019)

From an established CT assessment to interactive problem solving

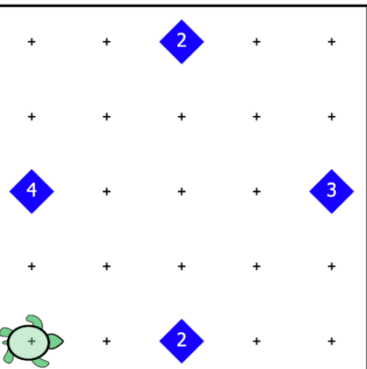
Student sample

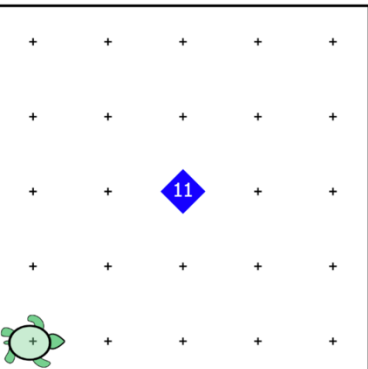
- 305 students from 4 countries engaging in 3 sessions of 45 min each.

Interactive problem-solving tasks

- 14 tasks, classified pre-analysis *"low complexity"* or *"high complexity"* based on a set of criteria (e.g. starter code, available computational elements, ...)

Challenge: Write code using while loops and more to help Karel solve the problem.

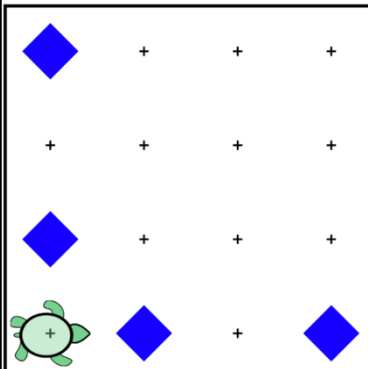
Start: 

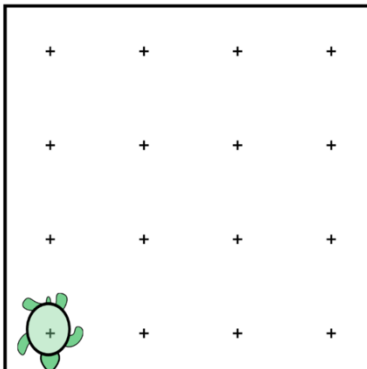
Goal: 

Code blocks: move forward, turn left, place stone, pickup stone, while front is clear, define main

play button, Play Speed (slow to fast)

Challenge: Place the pre-defined function in the code so Karel can solve the problem successfully.

Start: 

Goal: 

Code blocks: go back to start, pickup stones in line, move forward, turn left, place stone, pickup stone, define main, go back to start

play button, hint button, Play Speed (slow to fast)

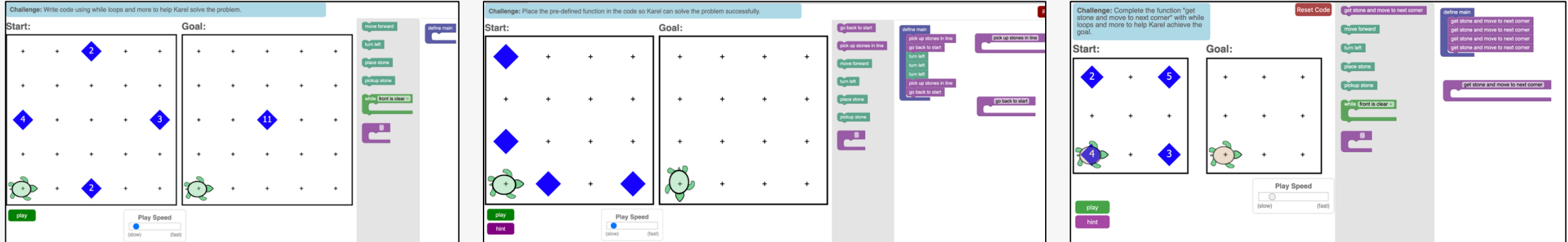
The relationship between B+CTt scores and CPS task success varied with task complexity

	Individual CPS Task Performance		
	Log-Odds	95% CI	p
(Intercept)	Lower complexity: $\beta = 1.42$, OR ≈ 4.1 Higher complexity: $\beta = .61$, OR ≈ 1.84		
Task Complexity (Low) B+CTt Score Prior Experience Task Complexity (Low): B+CTt Score			
Random Effects ICC Student:Teacher:Country ICC Teacher:Country ICC Country ICC Overall Adjusted			
Marginal R^2 / Conditional R^2	0.51 / 0.57		

Complexity impacts success rates and relationship between B+CTt and interactive performance.

3-level mixed-effects repeated-measures logistic regression on individual CPS task

What did our task-feature-based complexity distinction actually capture?



Task-feature variations impacting complexity matter – but at what level:

- **Difficulty:** How likely is the learner to succeed?
- **Competency demands / elicitation:** What knowledge or reasoning does the task require, and what opportunities does it create for learners to actually manifest it?
- **Evidentiary meaning:** What does the observed behavior tell us about competence?

Two complementary ways towards theorizing the task more richly for creating process evidence

Design levers → decision opportunities

Dimensions of decision opportunities:

Decision Allocation

Who decides: task or learner?

Available Options

Meaningful alternatives?

Decision Visibility

Need to decide explicit or implicit?



**Applied Computational Thinking
assessment framework.**

2026

Milestones as encoding task semantics

- Task decomposed into meaningful states that represent progress
- Process contextualized by the meaningful task state in which it occurs

OECD > Topics > Learning in the digital world > PISA 2025 Learning in the Digital World

**PISA 2025 Learning
in the Digital
World**

Task and tool design levers and decision opportunities

Instructions and prompts

Challenge: Place the pre-defined function in the code so Karel can solve the problem successfully.

Start:

Goal:

Problem structure

Starter Code

Available Code elements

Testing and feedback conditions

The screenshot displays the Karel programming environment. At the top, a challenge prompt asks the user to place pre-defined functions in the code to solve the problem. Below this, the 'Start' and 'Goal' states are shown on a 4x4 grid. The 'Start' state shows a turtle at (4,1) and stones at (1,1), (2,1), (3,1), (4,1), (1,2), (2,2), (3,2), (4,2), (1,3), (2,3), (3,3), (4,3), (1,4), (2,4), (3,4), (4,4). The 'Goal' state shows the same grid but with the stones moved to their target positions. The code editor on the right contains a 'define main' block with the following code: 'pick up stones in line', 'go back to start', 'turn left', 'turn left', 'pick up stones in line', 'go back to start'. The testing area at the bottom left has a 'play' button, a 'hint' button, and a 'Play Speed' slider.

Milestones: encoding task semantics as progress

Malik, ... & Piech (2019)

Milestone: a semantically meaningful task state or subgoal that marks progress toward a solution.

Task / problem structure

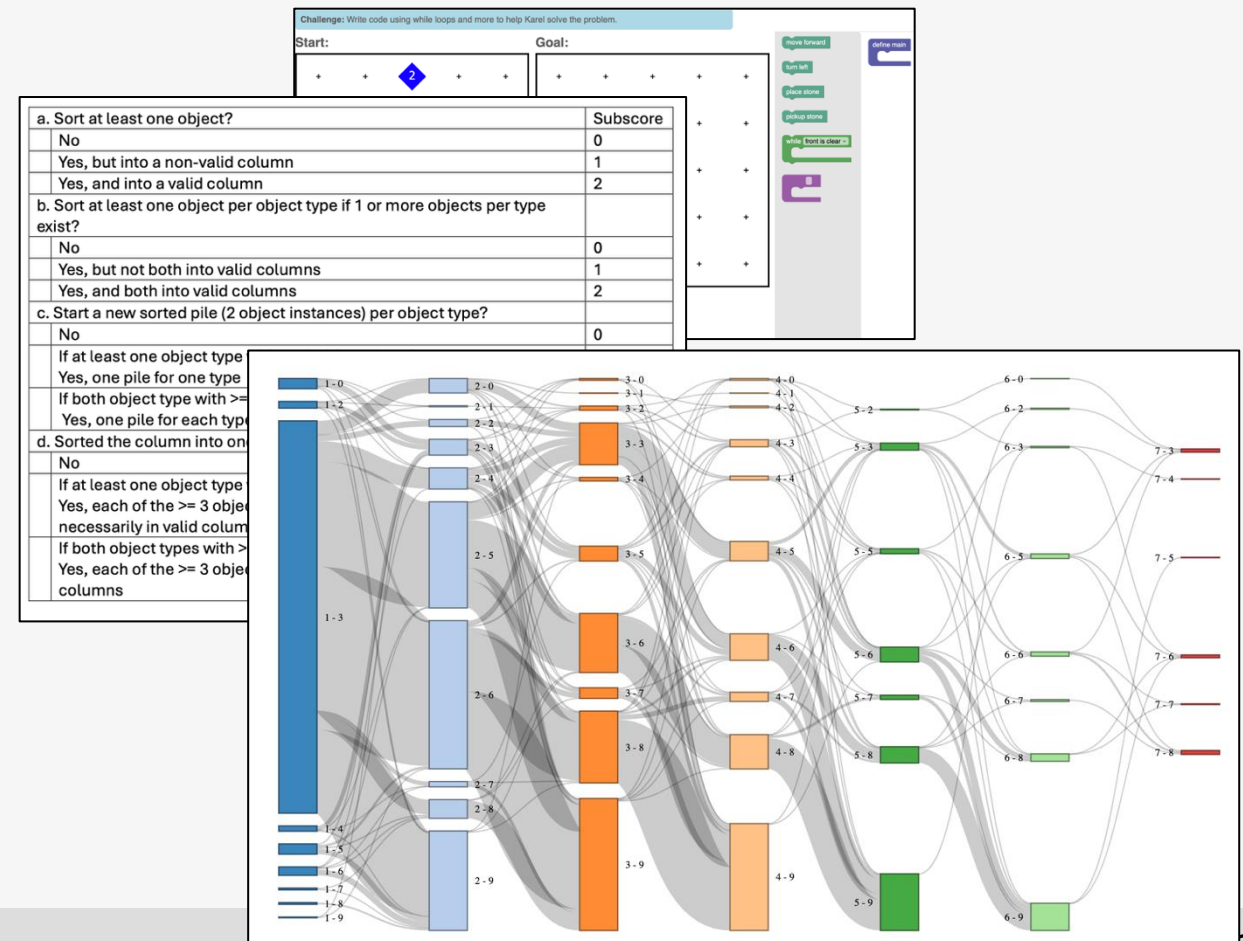


(computationally identifiable)

Milestones



Process interpreted relative to milestones reached and approached



What this work suggests

- Task variations can change what processes are elicited and what observed processes can, beyond decision opportunities.
- Knowing which task features matter is not enough; we need to richer theory to understand how design shapes the elicitation and interpretation of interaction process.

Structure of choice

Decision opportunities

*What does the learner get
to decide?*

Structure of progress

Milestones

*Where is the learner in the
problem?*

Structure of enactment

Interactional realization

*How is the decision
opportunity enacted in the
environment?*

Where next? Toward a predictive theory of task design for process evidence

- **Formalize:** How can we systematically represent decision opportunities, progress, and interactional realization?
- **Predict & validate:** Given a design change, can we predict how learner process will change, and test those predictions empirically?
- **Integrate with inference:** How should these task representations enter evidence models, so that process is interpreted in relation to the task conditions that produced it?

→ Automated Item Generation for interactive tasks



Thank you.

engin.bumbacher@hepl.ch

<https://www.linkedin.com/in/engin-bumbacher/>

Bumbacher, E., Salehi, S., Wieman, C., & Blikstein, P. (2018). Tools for science inquiry learning: Tool affordances, experimentation strategies, and conceptual understanding. *Journal of Science Education and Technology*, 27(3), 215-235.

Wang, K. D., Cock, J. M., Käser, T., & Bumbacher, E. (2023). A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. *British Journal of Educational Technology*, 54(1), 192-221.

Bumbacher, E., Brender, J., & Davis, R. L. (2024, March). A Comparative Analysis of Tools & Task Types for Measuring Computational Problem-Solving. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2* (pp. 1580-1581).