Assessing Outcomes and Processes of Student Collaboration

Peter F. Halpin

April 19, 2016

Joint work with: Alina von Davier, Yoav Bergner, Jiangang Hao, Lei Liu (ETS); Jacqueline Gutman (NYU)

NEW YORK UNIVERSITY

Outline

Part 1: Wherefore assessments involving collaboration?

- Set up the current perspective: performance assessments
- Selective review of research on small group productivity

Outline

Part 1: Wherefore assessments involving collaboration?

- Set up the current perspective: performance assessments
- Selective review of research on small group productivity

Part 2: Outcomes of collaboration

- Combining psychometric models with research on small group productivity
- Testing models against observed team performance

Outline

Part 1: Wherefore assessments involving collaboration?

- Set up the current perspective: performance assessments
- Selective review of research on small group productivity

Part 2: Outcomes of collaboration

- Combining psychometric models with research on small group productivity
- Testing models against observed team performance

Part 3: Processes of collaboration

- Focus on chat data (for now!)
- Modeling engagement among collaborators using temporal point processes[1]

[1] Halpin, von Davier, Hao, & Lui (under review). Journal of Educational Measurement.

Part 1: Why?

- 21st-century skills, non-cognitive skills, soft skills, hard-to-measure skills, social skills, ...
  - Theme: traditional educational tests target a relatively narrow set of constructs

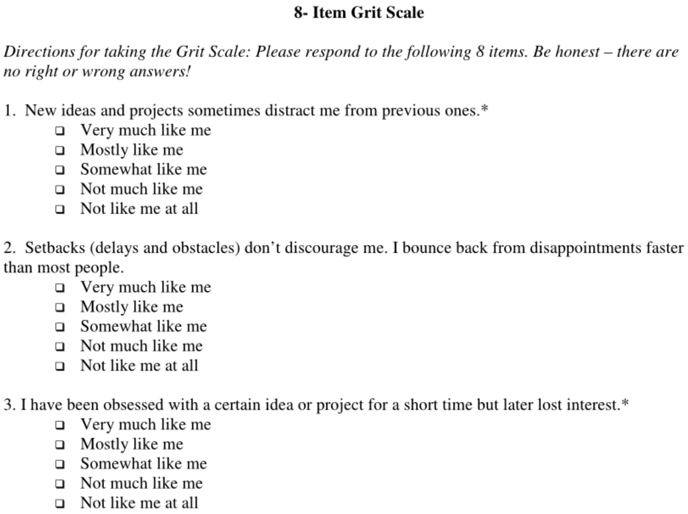
Part 1: Why?

- 21st-century skills, non-cognitive skills, soft skills, hard-to-measure skills, social skills, ...
  - Theme: traditional educational tests target a relatively narrow set of constructs
- Analyses of US labour markets indicate that such skills are valued by employers (Burrus et al.,2013; Deming, 2015)

Part 1: Why?

- 21st-century skills, non-cognitive skills, soft skills, hard-to-measure skills, social skills, ...
  - Theme: traditional educational tests target a relatively narrow set of constructs
- Analyses of US labour markets indicate that such skills are valued by employers (Burrus et al.,2013; Deming, 2015)
- There is a salient demand for assessments of a broader range of student competencies

With apologies to Dr. Duckworth...



8- Item Grit Scale

Directions for taking the Grit Scale: Please respond to the following 8 items. Be honest – there are no right or wrong answers!

1. New ideas and projects sometimes distract me from previous ones.*
   - □ Very much like me
   - □ Mostly like me
   - □ Somewhat like me
   - □ Not much like me
   - □ Not like me at all

2. Setbacks (delays and obstacles) don't discourage me. I bounce back from disappointments faster than most people.
   - □ Very much like me
   - □ Mostly like me
   - □ Somewhat like me
   - □ Not much like me
   - □ Not like me at all

3. I have been obsessed with a certain idea or project for a short time but later lost interest.*
   - □ Very much like me
   - □ Mostly like me
   - □ Somewhat like me
   - □ Not much like me
   - □ Not like me at all

`upenn.app.box.com/8itemgrit`

Self-reports

- Self-report measures often do not require the respondent to exhibit the skills about which we wish to make inferences

  → Unsuitable for supporting consequential decisions in educational settings[2]

[2] cf. Duckworth, & Yeager. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44(4)*, 237-251.

Educational assessments

- ☺ Reliability and generalizability in traditional content domains

Educational assessments

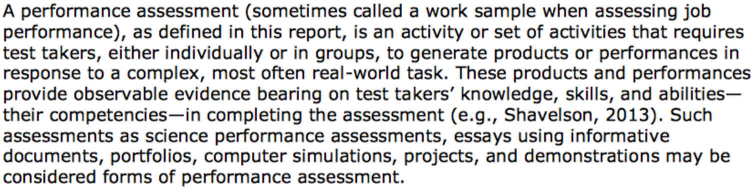
- ☺ Reliability and generalizability in traditional content domains
- ☹ Current psychometric models don't seem entirely appropriate to "next generation assessments"
  - e.g., IRT models don't use process data

Educational assessments

- ☺ Reliability and generalizability in traditional content domains
- ☹ Current psychometric models don't seem entirely appropriate to "next generation assessments"
  - e.g., IRT models don't use process data
- ☹ Collateral damage: teaching to the test, test anxiety, bubble-filling, ...
  - NY opt-out movement: 20% of students (parents) boycotted state test last year[3]

[3] www.wnyc.org/story/new-york-city-students-make-modest-gains-state-tests-opt-out-numbers-triple/

Performance assessments[4]



A performance assessment (sometimes called a work sample when assessing job performance), as defined in this report, is an activity or set of activities that requires test takers, either individually or in groups, to generate products or performances in response to a complex, most often real-world task. These products and performances provide observable evidence bearing on test takers' knowledge, skills, and abilities—their competencies—in completing the assessment (e.g., Shavelson, 2013). Such assessments as science performance assessments, essays using informative documents, portfolios, computer simulations, projects, and demonstrations may be considered forms of performance assessment.

[4] Davey, Ferrara, Holland, Shavelson, Webb, & Wise (2015). Psychometric Considerations for the Next Generation of Performance Assessment. Princeton, NJ. p. 10

Collaboration as a modality of performance assessment

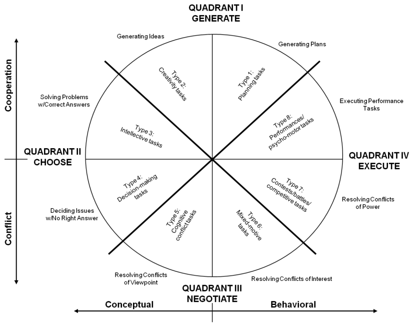
- Small group interactions are a highly-valued educational practice
  - The Jigsaw Classroom (Aronson et al., 1978; `jigsaw.org`)
  - Group-worthy tasks (Cohen et al., 1999)
- The use of information technology to support student collaboration is well established
  - CSCL (e.g., Hmelo-Silver et al., 2013)

Collaboration as a modality of performance assessment

- Small group interactions are a highly-valued educational practice
  - The Jigsaw Classroom (Aronson et al., 1978; `jigsaw.org`)
  - Group-worthy tasks (Cohen et al., 1999)
- The use of information technology to support student collaboration is well established
  - CSCL (e.g., Hmelo-Silver et al., 2013)
- The use of group work in assessment contexts has a relatively long-standing history
  - e.g., Webb, 1995; 2015

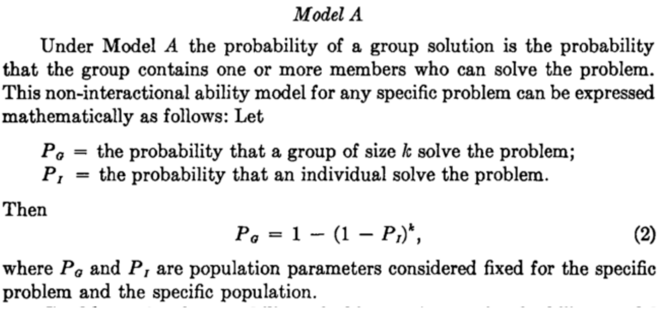
Intellective tasks

- Defined as having a demonstrably "correct" answer with respect to an agreed upon system of knowledge
- Differentiated from decision / judgement tasks on a continuum of *demonstrability* (Laughlin 2011)
- Differentiated from mixed-motive tasks in that the goals and outcomes are the same for all members


QUADRANT I
GENERATE
Generating Ideas
Generating Plans
Type 2: Creativity tasks
Type 1: Planning tasks
Cooperation
Solving Problems w/Correct Answers
Executing Performance Tasks
Type 3: Intellective tasks
Type 8: Performances/ psycho-motor tasks
QUADRANT II
CHOOSE
QUADRANT IV
EXECUTE
Type 4: Decision-making tasks
Type 7: Contests/battles/ competitive tasks
Resolving Conflicts of Power
Deciding Issues w/No Right Answer
Conflict
Type 5: Cognitive conflict tasks
Type 6: Mixed-motive tasks
Resolving Conflicts of Viewpoint
Resolving Conflicts of Interest
QUADRANT III
NEGOTIATE
Conceptual
Behavioral

McGrath's (1984) group task circumplex

Lorge & Solomon 1955[5]



Model A

Under Model A the probability of a group solution is the probability that the group contains one or more members who can solve the problem. This non-interactional ability model for any specific problem can be expressed mathematically as follows: Let

P_G = the probability that a group of size k solve the problem;
P_I = the probability that an individual solve the problem.

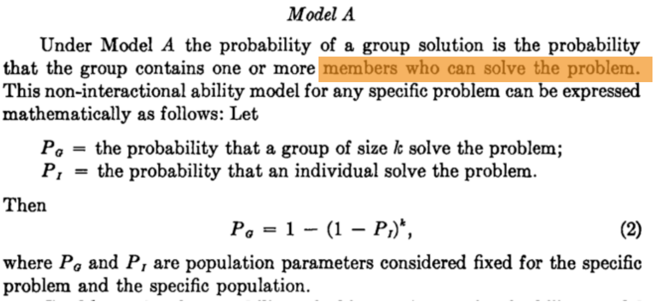
Then

$$P_G = 1 - (1 - P_I)^k, \qquad (2)$$

where P_G and P_I are population parameters considered fixed for the specific problem and the specific population.

[5] Two models of group behavior in the solution of Eureka-type problems. *Psychometrika, 1955, 20 (2)*, p. 141

Lorge & Solomon 1955[6]



Model A

Under Model *A* the probability of a group solution is the probability that the group contains one or more members who can solve the problem. This non-interactional ability model for any specific problem can be expressed mathematically as follows: Let

P_G = the probability that a group of size k solve the problem;
P_I = the probability that an individual solve the problem.

Then

$$P_G = 1 - (1 - P_I)^k, \qquad (2)$$

where P_G and P_I are population parameters considered fixed for the specific problem and the specific population.

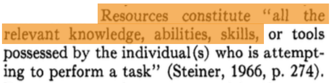
[6] Two models of group behavior in the solution of Eureka-type problems. *Psychometrika, 1955, 20 (2)*, p. 141

Smoke and Zajonc 1962[7]

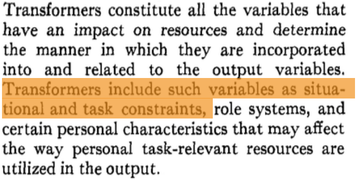
If p is the probability that a given individual member is correct, the group has a probability $h(p)$ of being correct, where $h(p)$ is a function of p depending upon the type of decision scheme accepted by the group. We shall call $h(p)$ a decision function. Intuitively, it would seem that a decision scheme is desirable to the extent that it surpasses p.

[7] On the reliability of group judgements and decisions. In *Mathematical methods for small group processes* (Eds. Criswell, Solomon, Suppes), p. 322

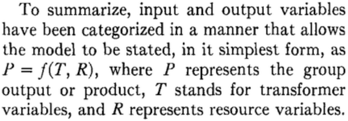
Schiflett 1979[8]



Resources constitute "all the relevant knowledge, abilities, skills, or tools possessed by the individual(s) who is attempting to perform a task" (Steiner, 1966, p. 274).



Transformers constitute all the variables that have an impact on resources and determine the manner in which they are incorporated into and related to the output variables. Transformers include such variables as situational and task constraints, role systems, and certain personal characteristics that may affect the way personal task-relevant resources are utilized in the output.



To summarize, input and output variables have been categorized in a manner that allows the model to be stated, in it simplest form, as $P = f(T, R)$, where P represents the group output or product, T stands for transformer variables, and R represents resource variables.

[8] Towards a general model of group productivity. *Psychological Bulletin, 86 (1), pp. 67-68*

Summary

- Building on research on small groups:
  - Intellective tasks (vs decision tasks)
  - Cooperative group interactions (vs competitive or mixed-motive)
  - Describing group outcomes via decision / functions that depend on characteristics of individuals
- But with a focus on:
  - Letting probability of success vary over individuals (e.g., via ability)
  - Describing relevant task characteristics (e.g., via difficulty)
  - The performance of individual groups rather than groups in aggregate

Outcomes of collaboration: A basic scenario

- Two students each write a conventional math assessment
- Their math ability is estimated to be θ_j and θ_k
- The two students then work together on a second conventional math assessment
- What do we expect about their performance on the second test, based on the first?

Collaboration as a psychometric question

- Traditional psychometric models assume conditional independence of the items

$$p(\mathbf{x}_j \mid \theta_j) = \prod_i^N p(x_{ij} \mid \theta_j) \tag{1}$$

- Traditional psychometric models also assume that the responses of two (or more) persons are independent

$$p(\mathbf{x}_j \; \mathbf{x}_k \mid \theta_j \; \theta_k) = p(\mathbf{x}_j \mid \theta_j) \; p(\mathbf{x}_k \mid \theta_k) \tag{2}$$

- When people work together does equation (2) hold?

"Working together" in terms of scoring rules[9]

- For binary items and pairs of responses, consider:
  - The conjunctive rule

$$x_{ijk} = \begin{cases} 1 & \text{if } x_{ij} = 1 \text{ and } x_{ik} = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - The disjunctive rule

$$x_{ijk} = \begin{cases} 0 & \text{if } x_{ij} = 0 \text{ and } x_{ik} = 0 \\ 1 & \text{otherwise} \end{cases}$$

- More possibilities, especially for items with > 2 responses or groups with > 2 collaborators

[9] cf. Steiner's 1966 classification of task types

Scoring rules vs decision functions

- Scoring rules describe what “counts” as a correct group response
    - Under control of the test designer[10]
- Decision functions describe the strategies adopted by a team
    - Under control of the team

[10] Maris & van der Maas (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika, 77 (4)*, 615-633

Scoring rules vs decision functions

- Scoring rules describe what “counts” as a correct group response
  - Under control of the test designer[10]
- Decision functions describe the strategies adopted by a team
  - Under control of the team
- Basic research strategy
  - Assume a certain scoring rule
  - Consider plausible models for team strategies
  - Test the models against data

[10] Maris & van der Maas (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika, 77 (4)*, 615-633

“Working together” in terms of scoring rules

- For binary items and pairs of responses, consider:
  - The conjunctive rule

$$x_{ijk} = \begin{cases} 1 & \text{if } x_{ij} = 1 \text{ and } x_{ik} = 1 \\ 0 & \text{otherwise} \end{cases}$$

  - The disjunctive rule

$$x_{ijk} = \begin{cases} 0 & \text{if } x_{ij} = 0 \text{ and } x_{ik} = 0 \\ 1 & \text{otherwise} \end{cases}$$

- More possibilities, especially for items with > 2 responses or groups with > 2 collaborators

Defining successful pairwise collaboration

- The independence model

$$E_{\text{ind}}[x_{ijk} \mid \theta_j \; \theta_k] = E[x_{ij} \mid \theta_j] \; E[x_{ik} \mid \theta_k]$$

Defining successful pairwise collaboration

- The independence model

$$E_{\text{ind}}[x_{ijk} \mid \theta_j \; \theta_k] = E[x_{ij} \mid \theta_j] \; E[x_{ik} \mid \theta_k]$$

- Successful collaboration

$$E[x_{ijk} \mid \theta_j \; \theta_k] > E_{\text{ind}}[x_{ijk} \mid \theta_j \; \theta_k]$$

- Unsuccessful collaboration

$$E[x_{ijk} \mid \theta_j \; \theta_k] < E_{\text{ind}}[x_{ijk} \mid \theta_j \; \theta_k]$$

- Note: these definitions are item- and dyad- specific

Some models for successful collaboration

- Minimum individual performance (disruptive team member)

$$E_{\min}[x_{ijk} \mid \theta_j \; \theta_k] = \min\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

Some models for successful collaboration

- Minimum individual performance (disruptive team member)

$$E_{\min}[x_{ijk} \mid \theta_j \; \theta_k] = \min\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

- Maximum individual performance (cheating / tutor)

$$E_{\max}[x_{ijk} \mid \theta_j \; \theta_k] = \max\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

Some models for successful collaboration

- Minimum individual performance (disruptive team member)

$$E_{\min}[x_{ijk} \mid \theta_j \; \theta_k] = \min\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

- Maximum individual performance (cheating / tutor)

$$E_{\max}[x_{ijk} \mid \theta_j \; \theta_k] = \max\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

- “True collaboration”

$$E[x_{ijk} \mid \theta_j \; \theta_k] \geq \max\{E[x_{ij} \mid \theta_j], E[x_{ik} \mid \theta_k]\}$$

A model for “true collaboration”

- An additive model

$$E_{\text{add}}[x_{ijk} \mid \theta_j \ \theta_k] = E[x_{ij} \mid \theta_j] + E[x_{ik} \mid \theta_k] - E[x_{ijk} \mid \theta_j \ \theta_k]$$

A model for "true collaboration"

- An additive model

$$E_{\mathrm{add}}[x_{ijk} \mid \theta_j \; \theta_k] = E[x_{ij} \mid \theta_j] + E[x_{ik} \mid \theta_k] - E[x_{ijk} \mid \theta_j \; \theta_k]$$

- Recalling $E[x_{ijk} \mid \theta_j \; \theta_k] > E[x_{ij} \mid \theta_j] E[x_{ik} \mid \theta_k]$, define an additive independence (AI) model

$$E_{AI}[x_{ijk} \mid \theta_j \; \theta_k] = E[x_{ij} \mid \theta_j] + E[x_{ik} \mid \theta_k] - E[x_{ij} \mid \theta_j] \, E[x_{ik} \mid \theta_k]$$

$$\geq E_{\mathrm{add}}[x_{ijk} \mid \theta_j \; \theta_k]$$

- AI is an upper bound on any "more interesting" additive model for successful collaboration

More on AI model

- Can also be written as:

$$\begin{aligned} E_{AI}[x_{ijk} \mid \theta_j \; \theta_k] = {} & E[x_{ij} \mid \theta_j]\,(1 - E[x_{ik} \mid \theta_k]) \\ & + E[x_{ik} \mid \theta_k]\,(1 - E[x_{ij} \mid \theta_j]) \\ & + E[x_{ij} \mid \theta_j]\,E[x_{ik} \mid \theta_k] \end{aligned}$$

- Which has an interpretation in terms of three cases

More on AI model

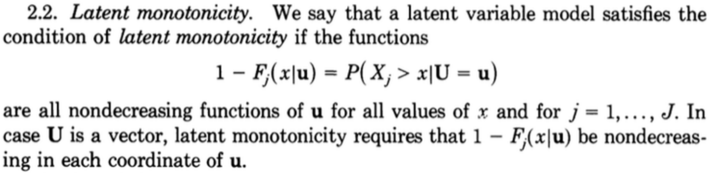
- And is also equivalent to Lorge & Solomon's Model A

$$E_{AI}[x_{ijk} \mid \theta_j \; \theta_k] = 1 - (1 - E[x_{ij} \mid \theta_j])(1 - E[x_{ik} \mid \theta_k])$$

- Except the "probability an individual can solve the problem" now depends on both the individual and the problem

More on AI model[11]

- We probably want some constraints on what counts as a good collaborative IRF



2.2. *Latent monotonicity.* We say that a latent variable model satisfies the condition of *latent monotonicity* if the functions

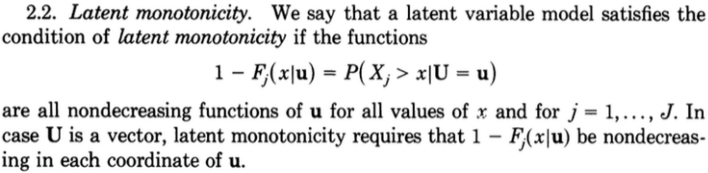
$$1 - F_j(x|\mathbf{u}) = P(X_j > x|\mathbf{U} = \mathbf{u})$$

are all nondecreasing functions of $\mathbf{u}$ for all values of x and for $j = 1, \ldots, J$. In case $\mathbf{U}$ is a vector, latent monotonicity requires that $1 - F_j(x|\mathbf{u})$ be nondecreasing in each coordinate of $\mathbf{u}$.

[11] Holland & Rosenbaum (1986). Conditional Association and Unidimensionality in Monotone Latent Variable Models. *The Annals of Statistics, 14 (4)*, 1523 – 1543

More on AI model[11]

- We probably want some constraints on what counts as a good collaborative IRF



2.2. *Latent monotonicity*. We say that a latent variable model satisfies the condition of *latent monotonicity* if the functions

$$1 - F_j(x|\mathbf{u}) = P(X_j > x|\mathbf{U} = \mathbf{u})$$

are all nondecreasing functions of **u** for all values of x and for $j = 1, \ldots, J$. In case **U** is a vector, latent monotonicity requires that $1 - F_j(x|\mathbf{u})$ be nondecreasing in each coordinate of **u**.

- Easy to show that AI satisfies latent monotonicity, if the individual IRFs do (trivial for other models also)

[11] Holland & Rosenbaum (1986). Conditional Association and Unidimensionality in Monotone Latent Variable Models. *The Annals of Statistics, 14 (4)*, 1523 – 1543

AI: latent monotonicity

Assumptions:

$$f(x) \geq f(x') \text{ for } x > x' \quad \text{and} \quad 0 \leq g(y) \leq 1$$

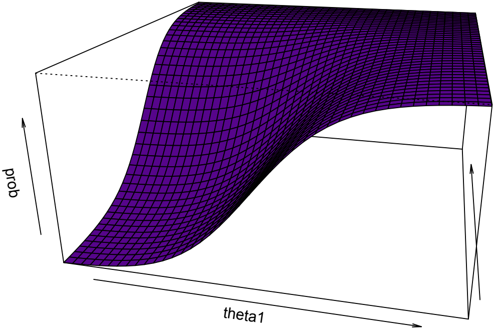
Show:

$$f(x) + g(y) - f(x)\,g(y) \geq f(x') + g(y) - f(x')\,g(y)$$

Contradiction:

$$\begin{gathered} f(x) + g(y) - f(x)\,g(y) < f(x') + g(y) - f(x')\,g(y) \\ \rightarrow f(x) - f(x') < g(y)\,(f(x) - f(x')) \end{gathered}$$

AI: example IRF[12]



[12] Using 2PL model for individual IRFs with $\alpha = 1$ and $\beta = 0$

Models abound!

- Basic idea: write down IRFs for collaboration based on assumed-to-be-known individual abilities (and item parameters)
- But how do we characterize empirical team performance?

Empirical team performance

- We have
  - Observed collaborative responses $\mathbf{x}_{jk} = (x_{1jk}, x_{1jk}, \ldots, x_{mjk})$
  - A model for individual performance on the m (conventional) math items

Empirical team performance

- We have
  - Observed collaborative responses $\mathbf{x}_{jk} = (x_{1jk}, x_{1jk}, \ldots, x_{mjk})$
  - A model for individual performance on the m (conventional) math items
- So we can get "team theta," e.g.,

$$\hat{\theta}_{jk} = \underset{\theta}{\operatorname{argmax}} \, \{L_0(\mathbf{x}_{jk} \mid \theta)\} \tag{3}$$

- Where L_0 is the likelihood of the model calibrated on individual performance (reference model)

Proposed method for testing models

- Testing of different models against reference model

$$D_{\text{model}} = -2\ln \frac{L_{\text{model}}(\mathbf{x}_{jk} \mid \theta_j\ \theta_k)}{L_0(\mathbf{x}_{jk} \mid \hat{\theta}_{jk})} \tag{4}$$

- Also a “direct test” of effect of collaboration for each individual

$$D_0 = -2\ln \frac{L_0(\mathbf{x}_{jk} \mid \theta_j)}{L_0(\mathbf{x}_{jk} \mid \hat{\theta}_{jk})} \tag{5}$$

with effect size $\delta_{jk} = \frac{\theta_{jk}-\theta_j}{\sigma_\theta}$

Proposed method: reference distribution

- Ind and AI models are not nested with reference model → No Wilk's theorem
- Can use Vuong's 1989[13] results for LR with non-nested models, but asymptotic in m
- Good news: we can bootstrap a null distribution for (4) and (5) pretty easily

[13] Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrika, 57(2)*, 307 – 333.

Bootstrapping the reference distribution

Assuming known item parameters and θ_j, θ_k. For $r = 1, \ldots, R$

Step 1 Generate collaborative response patterns $\mathbf{x}_{jk}^{(r)}$ from $E_{\mathrm{model}}[x_{ijk} \mid \theta_j \; \theta_k]$

Step 2 Compute $L_{\mathrm{model}}(\mathbf{x}_{jk}^{(r)} \mid \theta_j \; \theta_k)$

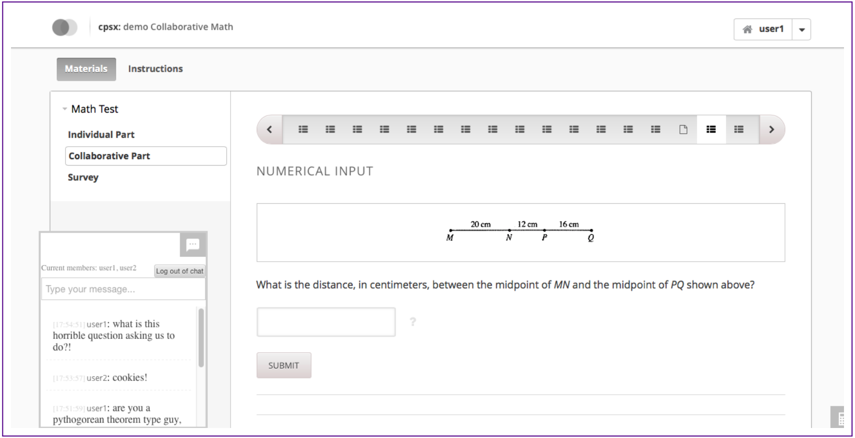
Step 2 Estimate $\theta_{jk}^{(r)}$ for each $\mathbf{x}_{jk}^{(r)}$; save $L_0(\mathbf{x}_{jk}^{(r)} \mid \theta_{jk})$

Step 4 Compute $D_{\mathrm{model}}^{(r)}$ or $D_0^{(r)}$

Example 1

- Design
  - Pool of pre-calibrated math items (grade 12 NAEP, modified to be numeric response)
  - Individual "pre-test" → estimate individual abilities
  - Collaborative "post-test" → evaluate models, estimate δ_{jk}
  - Modality of collaboration: online chat
- Limitations:
  - Small calibration sample; crowd workers
  - Individual and collaborative forms were not counterbalanced (neither in order nor content)

NAEP grade 12 math items, deployed via OpenEdx



What is the distance, in centimeters, between the midpoint of *MN* and the midpoint of *PQ* shown above?

Current members: user1, user2

[17:54:51] user1: what is this horrible question asking us to do?!

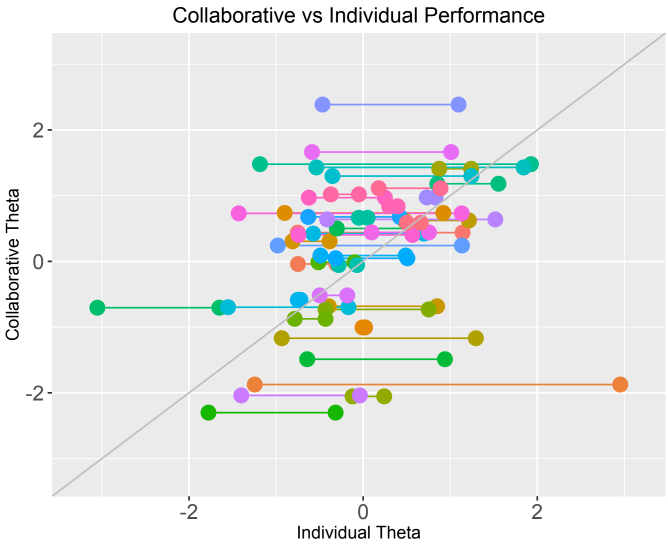
[17:53:57] user2: cookies!

[17:51:59] user1: are you a pythogorean theorem type guy,

AMT crowdworkers (calibration sample)

Variable	**Levels**	**n**	%	$\sum$ %
Gender	Female	155	46.5	46.5
	Male	178	53.5	100.0
Age	18-30	117	35.2	35.2
	30-40	129	38.9	74.1
	40-55	71	21.4	95.5
	55+	15	4.5	100.0
Education	Some Grade School	3	0.9	0.9
	High School Diploma	49	14.7	15.6
	Some College	118	35.4	51.0
	Bachelor's Degree	132	39.6	90.7
	Master's Degree	22	6.6	97.3
	Ph.D or Advanced Degree	9	2.7	100.0
Country	United States	313	94.0	94.0
	India	16	4.8	98.8
	Canada	3	0.9	99.7
	United Kingdom	1	0.3	100.0
English First Lang	Yes	321	96.4	96.4
	No	12	3.6	100.0

Deltas


Collaborative vs Individual Performance
Collaborative Theta
Individual Theta
2
0
-2
-2
0
2

Model tests: Sanity check using individual pre-test

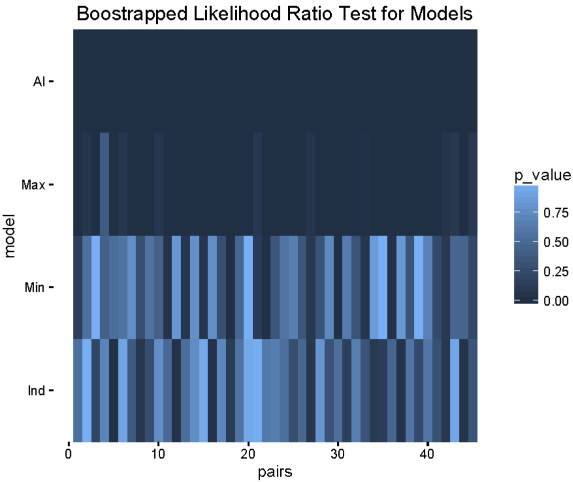


Figure reports P($D_{\text{model}} > |obs|$) for individual pre-tests scored using conjunctive scoring rule

Model tests: Collaborative data

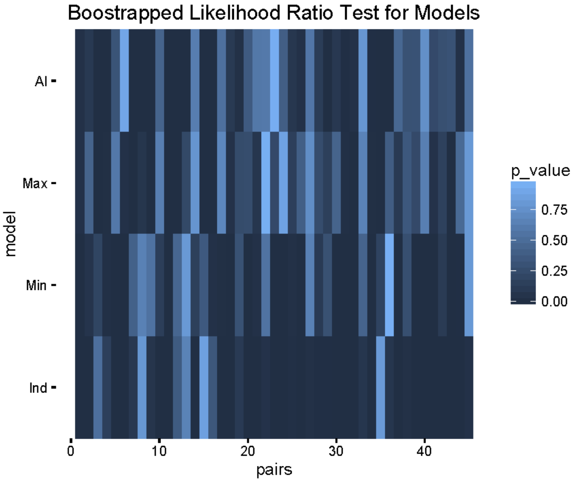
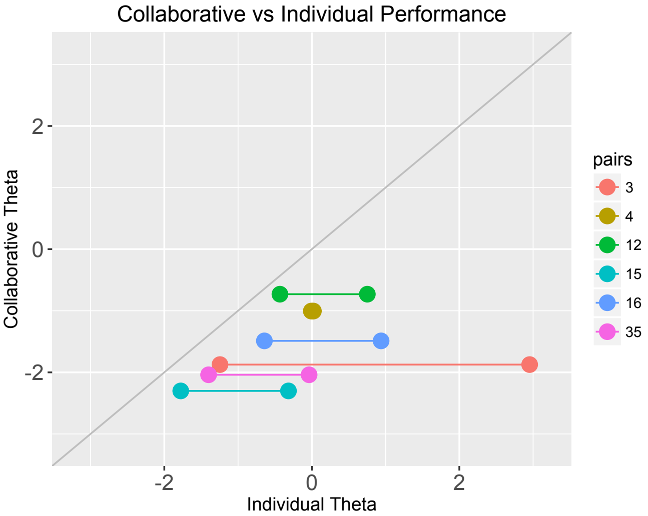
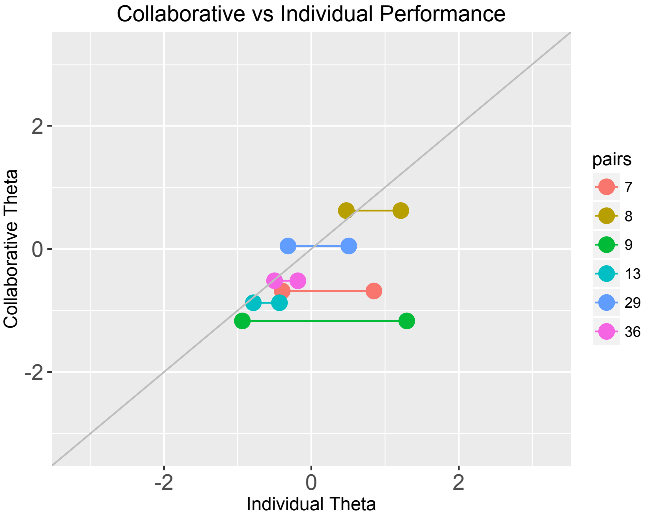


Figure reports P$(D_{\text{model}} > |obs|)$ for collaborative tests scored using conjunctive scoring rule

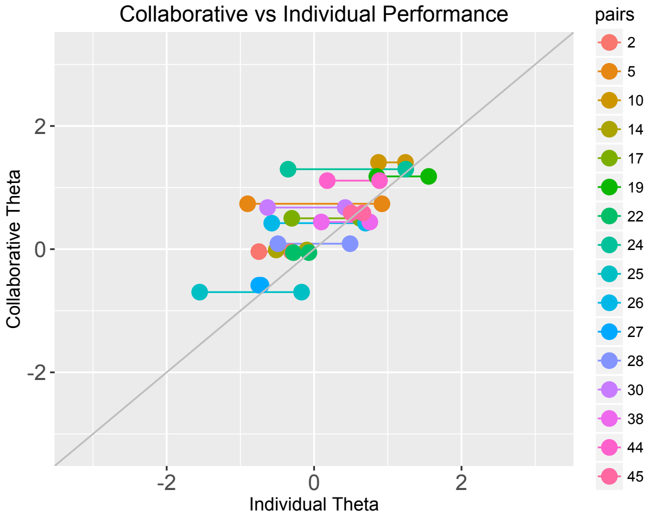
Ind Model


Collaborative vs Individual Performance
Collaborative Theta
Individual Theta
2
0
-2
-2
0
2
pairs
3
4
12
15
16
35

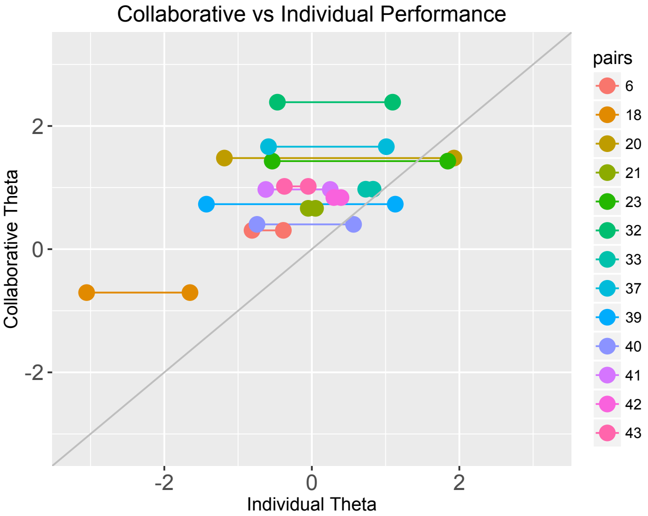
Min Model


Collaborative vs Individual Performance
Collaborative Theta
Individual Theta
2
0
-2
-2
0
2
pairs
7
8
9
13
29
36

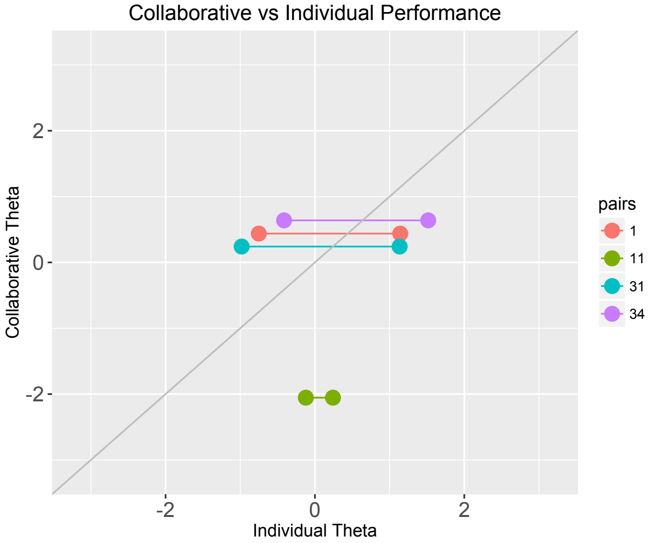
Max Model


Collaborative vs Individual Performance
Collaborative Theta
Individual Theta
2
0
-2
-2
0
2
pairs
2
5
10
14
17
19
22
24
25
26
27
28
30
38
44
45

AI Model


Collaborative vs Individual Performance
Collaborative Theta
Individual Theta
2
0
-2
-2
0
2
pairs
6
18
20
21
23
32
33
37
39
40
41
42
43

Not one of our four models



Summary of collaborative outcomes

- Can define, estimate, and test models of collaboration on academic performance using IRT-based methods
- But how distinct are these models, really?
- Models do not cover all cases

Possible next step – one model to rule them all!

Let $w_1, w_2 \in [0, 1]$ and define the weighted additive independence model

$$E_{\text{WAI}}[X_{ijk} \mid \theta_j\ \theta_k] = w_j P_i(\theta_j)\, Q_i(\theta_k) + w_k P_i(\theta_k)\, Q_i(\theta_j) + P_i(\theta_j)\, P_i(\theta_k)$$

- Includes original four and everything in between
- Includes $(P_i(\theta_j) + P_i(\theta_k))/2$ when $w_1 = w_2 = .5$
- Weights describe how well each individual obtains his/her “optimal collaboration level”

Part 3: What are process data?[14]

[14] Halpin & von Davier 2013, Hao, & Lui (under review). Journal of Educational Measurement.

Part 3: What are process data?[14]

- Any task-related actions of a respondent performed during the completion of a task
  - In ed tech context, typically associated with time-stamped user logs ("trace data")

[14] Halpin & von Davier 2013, Hao, & Lui (under review). Journal of Educational Measurement.

Part 3: What are process data?[14]

- Any task-related actions of a respondent performed during the completion of a task
  - In ed tech context, typically associated with time-stamped user logs ("trace data")
- All the stuff IRT ignores:

$$p(\mathbf{x} \mid \theta) = \prod_i p(x_i \mid \theta)$$

[14] Halpin & von Davier 2013, Hao, & Lui (under review). Journal of Educational Measurement.

Part 3: What are collaborative process data?

- Ideally a richly detailed recording of the sequence of actions taken by each team member during the completion of a task
  - ATC21S collaborative problem solving prototype items[15]
  - CPS frame[16]

[15] http://www.atc21s.org/uploads/3/7/0/0/37007163/pd_module_3_nonadmin.pdf

[16] In alpha at Computational Psychometrics lab at ETS

Part 3: What are collaborative process data?

- Ideally a richly detailed recording of the sequence of actions taken by each team member during the completion of a task
  - ATC21S collaborative problem solving prototype items[15]
  - CPS frame[16]
- Focus today: chat messages sent between online collaborators

[15] http://www.atc21s.org/uploads/3/7/0/0/37007163/pd_module_3_nonadmin.pdf

[16] In alpha at Computational Psychometrics lab at ETS

Two perspectives on the analysis of chat / email / etc.

- Text-based analysis of strategy and sentiment
  - e.g., Howley, Mayfield, & Rosè, 2013; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015
- Time series analysis of sending times
  - e.g., Barabàsi, 2005; Ebel, Mielsch, & Bornholdt, 2002; Halpin & De Boeck, 2013

Temporal point process: basic idea (more tomorrow)[17]

- Data: events that have negligible duration relative to a period of observation
  - Contrast events with states, regimes
- Basic idea: model the Bernoulli probability of an event happening in a small window of time $[t, t + \Delta)$, conditional on the events that have happened before $t \in \mathbb{R}+$.
  - "Instantaneous probability" of an event, denoted $p(t)$

[17] Daley, D. J., & Vera-Jones. (2003). An introduction to the theory of point processes: Elementary theory and methods (2nd ed., Vol. 1). New York: Springer.

Temporal point process in interpersonal context

- Modeling $p(t)$ to describe
    - How the probability of each person's actions changes in continuous time
    - How this depends on their previous actions
    - Emergent or group-level phenomena like coordination, reciprocity, ...

Chat engagement via Hawkes processes

- Hawkes process provides a means of modeling instantaneous probabilities in a multivariate context
- Halpin et al. (under review) suggest the response intensity parameter as a measure of engagement of student j with k

$$\alpha_{jk} > \bar{n}_{jk}/n_k \tag{6}$$

  - $\bar{n}_{jk}$ is the expected total number of responses made by student j to student k is (inferred from model)
  - n_k is the number of actions of student k (observed)
  - Lower bound is tight in practice; not necessary for computations

Chat engagement via Hawkes processes

- Aggregating to team (dyad) level

$$\alpha \equiv \frac{\alpha_{12} n_2 + \alpha_{21} n_1}{n_1 + n_2} \tag{7}$$

- Interpretation: the proportion of all group members' actions, $n_1 + n_2$, that were responded to by any other member during a collaboration

- See paper for more details, including initial results on SEs of α_{jk}

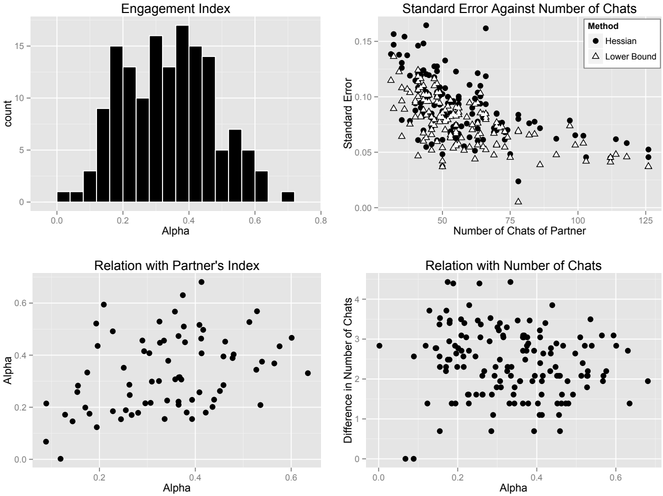
Example: Tetralogue

- A simulation-based science game with an embedded assessment recently developed at ETS (Hao, Liu, von Davier, & Kyllonen, 2015)
    1. Dyads work together to learn and make predictions about volcano activity
    2. At various points in the simulation, the students are asked to individually submit their responses to an assessment item without discussing the item
    3. Following submission of responses from both students, they are invited to discuss the question and their answers
    4. Lastly, they are given an opportunity to revise their responses to the item, with the final answers counting towards the team's score

Example: Full sample

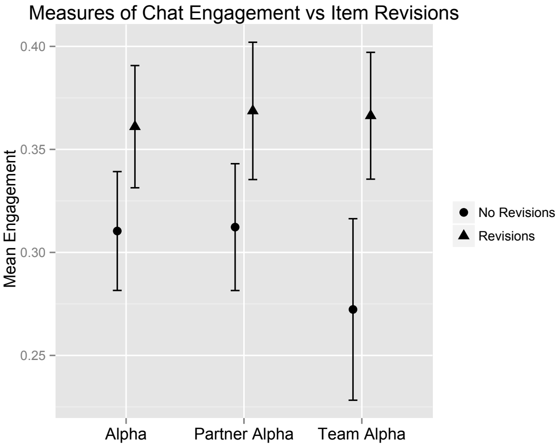
- 286 dyads solicited via AMT and randomly paired (based on arrival in queue)
- Median reported age was 31.5 years
- 52.5% reported that they were female
- 79.2% reported that they were White.
- Additionally, all participants were required to
  - Have an IP address located in the United States
  - Self-identify as speaking English as their primary language
  - Self-identify as having at least one year of college education

Example: Estimating chat engagement



Note: Alpha denotes the estimated response intensities from Equation 6. Hessian denotes standard errors obtained via the Hessian of the log-likelihood. See appendix of Halpin et al. for Lower Bound. Difference in Number of Chats was scaled using the log of the absolute value of the difference.

Example: Relation with revision on embedded assessment



Note: Comparison of mean levels of engagement indices for individuals who either did or did not revise at least one response after discussion with their partners. Alpha denotes the estimated response intensities from Equation 6; Partner's Alpha denotes the partner's response intensity; Team Alpha denotes the team-level index in Equation 7. For the latter, the data are reported for dyads, not individuals, and no revisions means that both individuals on the team made no revisions. Error bars are 95% confidence intervals on the means.

Example: Relation with revision on embedded assessment

Table 1: Summary of group differences.

Index	Group	Mean	SD	N	Hedges' g	r
Alpha	No Revisions	0.31	0.13	82	–	
Alpha	Revisions	0.36	0.10	66	0.40	.20
Partner's Alpha	No Revisions	0.31	0.14	82	–	
Partner's Alpha	Revisions	0.37	0.14	66	0.44	.21
Team Alpha	No Revisions	0.27	0.11	26	–	
Team Alpha	Revisions	0.37	0.13	48	0.84	.38

Note: Alpha denotes the estimated response intensities from Equation alpha2; Partner's Alpha denotes the engagement index of the individual's partner; Team Alpha denotes the team-level index in Equation 7. Hedges' g used the correction factor described by Hedges (1981) and r denotes the point-biserial correlation.

Summary of collaborative processes

- Hawkes processes are a feasible model for process data obtained on collaborative tasks
- Resulting measures of chat engagement are meaningfully related to task performance
- Future modeling work
    - Random effects models for simultaneous estimation over multiple groups
    - Inclusion of model parameters describing task characteristics
    - Analytic expressions for standard errors of model parameters
    - Methods for improving optimization with relatively small numbers of events
    - Integration with text-based analyses (e.g., using marks / time-varying covariates)

What's next

- Integration of task design, outcomes, processes, ... and theory!!

Contact: peter.halpin@nyu.edu

Support: This research was funded by a postdoctoral fellowship from the Spencer Foundation and an Education Technology grant from NYU Steinhardt.

References not already included in footnotes

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). The jigsaw classroom. Beverly Hills, CA: Sage.

Burrus, J., Carlson, J., Bridgeman, B., Golub-smith, M., & Greenwood, R. (2013). Identifying the Most Important 21st Century Workforce Competencies : An Analysis of the Occupational Information Network (O * NET) (ETS RR-13-21). Princeton, NJ.

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. Theory Into Practice, 38, 80-86.

Davey, T., Ferrara, S., Holland, P. W., Shavelson, R. J., Webb, N. M., & Wise, L. L. (2015). Psychometric Considerations for the Next Generation of Performance Assessment. Princeton, NJ.

Deming, D. J. (2015). The Growing Importance of Social Skills in the Labor Market. National Bureau of Economic Research Working Paper Series, (21473).

Griffin, P., & Care, E. (2015). Assessment and teaching of 21st century skills: Methods and approach. New York: Springer.

Hmelo-Silver, C. E., Chinn, C. A., Chan, C. K., & O?Donnel, A. M. (2013). International handbook of collaborative learning. New York: Taylor and Francis.

McGrath, J. E. (1984). Groups: Interaction and performance. (Prentice-Hall, Ed.). Englewood Cliffs, NJ.

Organisation for Economic Co-operation and Development. (2013). PISA 2015 Draft Collaborative Problem Solving Framework. Retrieved from `http://www.oecd.org/pisa/pisaproducts/DraftPISA2015CollaborativeProblemSolvingFramework.pdf`

Webb, N. M. (1995). Group Collaboration in Assessment: Multiple Objectives, Processes, and Outcomes. Educational Evaluation and Policy Analysis, 17(2), 239-261.

Bootstrapping the reference distribution for D_{model}

Assuming known item parameters and θ_j, θ_k. For $r = 1, \ldots, R$

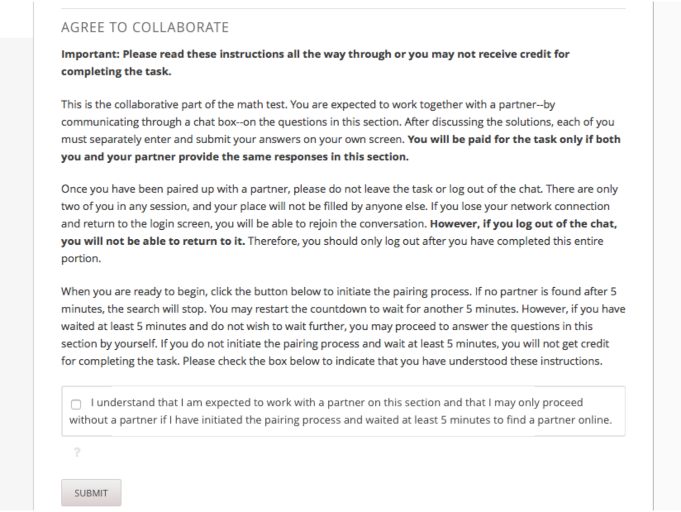
Step 1 Generate collaborative response patterns $\mathbf{x}_{jk}^{(r)}$ from $E_{\mathrm{model}}[x_{ijk} \mid \theta_j \ \theta_k]$

Step 2 Compute $L_{\mathrm{model}}(\mathbf{x}_{jk}^{(r)} \mid \theta_j \ \theta_k)$

Step 2 Estimate $\theta_{jk}^{(r)}$ for each $\mathbf{x}_{jk}^{(r)}$; save $L_0(\mathbf{x}_{jk}^{(r)} \mid \theta_{jk})$

Step 4 Compute $D_{\mathrm{model}}^{(r)}$ or $D_0^{(r)}$

Instructions



AGREE TO COLLABORATE

Important: Please read these instructions all the way through or you may not receive credit for completing the task.

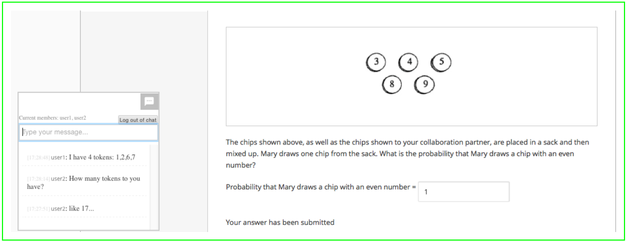
This is the collaborative part of the math test. You are expected to work together with a partner--by communicating through a chat box--on the questions in this section. After discussing the solutions, each of you must separately enter and submit your answers on your own screen. **You will be paid for the task only if both you and your partner provide the same responses in this section.**

Once you have been paired up with a partner, please do not leave the task or log out of the chat. There are only two of you in any session, and your place will not be filled by anyone else. If you lose your network connection and return to the login screen, you will be able to rejoin the conversation. **However, if you log out of the chat, you will not be able to return to it.** Therefore, you should only log out after you have completed this entire portion.

When you are ready to begin, click the button below to initiate the pairing process. If no partner is found after 5 minutes, the search will stop. You may restart the countdown to wait for another 5 minutes. However, if you have waited at least 5 minutes and do not wish to wait further, you may proceed to answer the questions in this section by yourself. If you do not initiate the pairing process and wait at least 5 minutes, you will not get credit for completing the task. Please check the box below to indicate that you have understood these instructions.

□ I understand that I am expected to work with a partner on this section and that I may only proceed without a partner if I have initiated the pairing process and waited at least 5 minutes to find a partner online.

Jigsaw / information sharing items



[17:28:48] user1: I have 4 tokens: 1,2,6,7

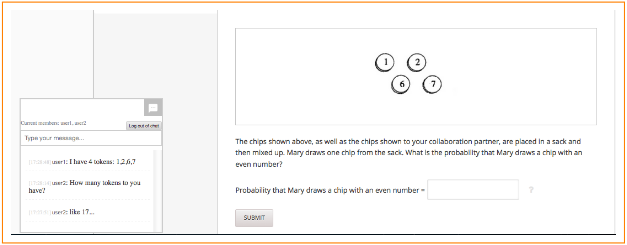
[17:28:14] user2: How many tokens to you have?

[17:27:51] user2: like 17...

The chips shown above, as well as the chips shown to your collaboration partner, are placed in a sack and then mixed up. Mary draws one chip from the sack. What is the probability that Mary draws a chip with an even number?

Probability that Mary draws a chip with an even number = 1

Your answer has been submitted



[17:28:48] user1: I have 4 tokens: 1,2,6,7

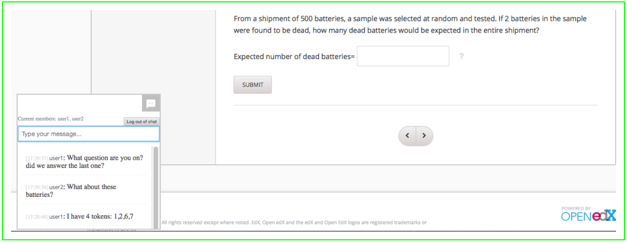
[17:28:14] user2: How many tokens to you have?

[17:27:51] user2: like 17...

The chips shown above, as well as the chips shown to your collaboration partner, are placed in a sack and then mixed up. Mary draws one chip from the sack. What is the probability that Mary draws a chip with an even number?

Probability that Mary draws a chip with an even number =

Jigsaw / information sharing items



From a shipment of 500 batteries, a sample was selected at random and tested. If 2 batteries in the sample were found to be dead, how many dead batteries would be expected in the entire shipment?

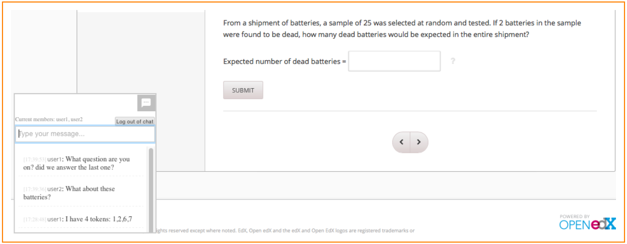
Expected number of dead batteries=

[17:39:53] user1: What question are you on? did we answer the last one?

[17:39:36] user2: What about these batteries?

[17:28:48] user1: I have 4 tokens: 1,2,6,7

All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or



From a shipment of batteries, a sample of 25 was selected at random and tested. If 2 batteries in the sample were found to be dead, how many dead batteries would be expected in the entire shipment?

Expected number of dead batteries =

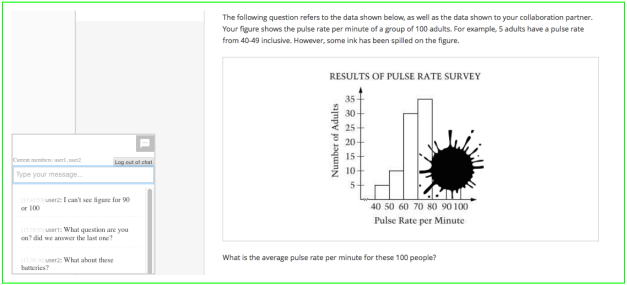
[17:39:53] user1: What question are you on? did we answer the last one?

[17:39:36] user2: What about these batteries?

[17:28:48] user1: I have 4 tokens: 1,2,6,7

ights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or

Jigsaw / information sharing items



Current members: user1 , user2

Type your message...

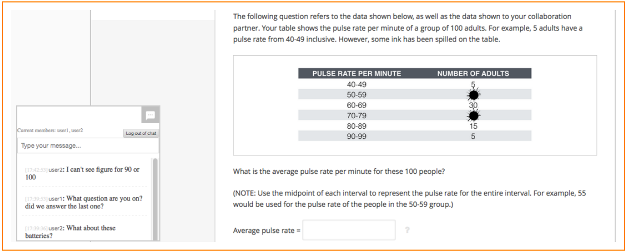
[17:42:53] user2: I can't see figure for 90 or 100

[17:39:53] user1: What question are you on? did we answer the last one?

[17:39:36] user2: What about these batteries?

The following question refers to the data shown below, as well as the data shown to your collaboration partner. Your figure shows the pulse rate per minute of a group of 100 adults. For example, 5 adults have a pulse rate from 40-49 inclusive. However, some ink has been spilled on the figure.

What is the average pulse rate per minute for these 100 people?



Current members: user1 , user2

Type your message...

[17:42:53] user2: I can't see figure for 90 or 100

[17:39:53] user1: What question are you on? did we answer the last one?

[17:39:36] user2: What about these batteries?

The following question refers to the data shown below, as well as the data shown to your collaboration partner. Your table shows the pulse rate per minute of a group of 100 adults. For example, 5 adults have a pulse rate from 40-49 inclusive. However, some ink has been spilled on the table.

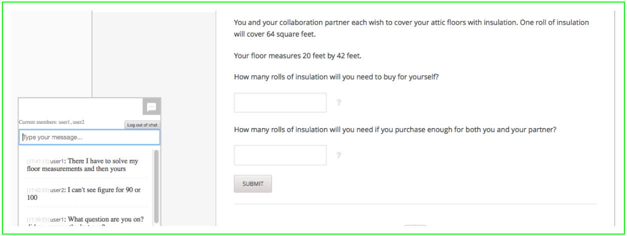
PULSE RATE PER MINUTE	NUMBER OF ADULTS
40-49	5
50-59	
60-69	30
70-79	
80-89	15
90-99	5

What is the average pulse rate per minute for these 100 people?

(NOTE: Use the midpoint of each interval to represent the pulse rate for the entire interval. For example, 55 would be used for the pulse rate of the people in the 50-59 group.)

Average pulse rate =

Jigsaw / information sharing items



Current members: user1, user2

[17:47:13] user1: There I have to solve my floor measurements and then yours

[17:42:53] user2: I can't see figure for 90 or 100

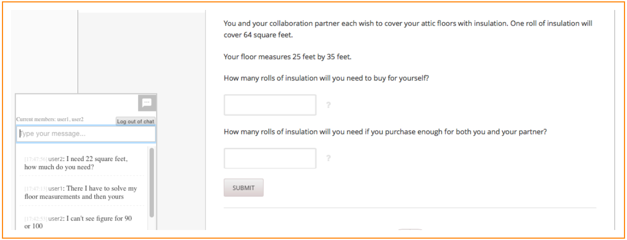
[17:39:53] user1: What question are you on?

You and your collaboration partner each wish to cover your attic floors with insulation. One roll of insulation will cover 64 square feet.

Your floor measures 20 feet by 42 feet.

How many rolls of insulation will you need to buy for yourself?

How many rolls of insulation will you need if you purchase enough for both you and your partner?



Current members: user1, user2

[17:47:56] user2: I need 22 square feet, how much do you need?

[17:47:13] user1: There I have to solve my floor measurements and then yours

[17:42:53] user2: I can't see figure for 90 or 100

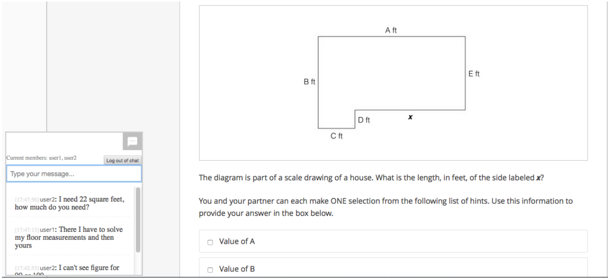
You and your collaboration partner each wish to cover your attic floors with insulation. One roll of insulation will cover 64 square feet.

Your floor measures 25 feet by 35 feet.

How many rolls of insulation will you need to buy for yourself?

How many rolls of insulation will you need if you purchase enough for both you and your partner?

Hints / information requesting items



Current members: user1, user2

Type your message...

[17:47:56] user2: I need 22 square feet, how much do you need?

[17:47:13] user1: There I have to solve my floor measurements and then yours

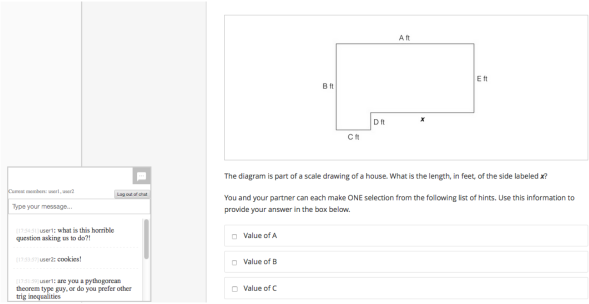
[17:42:53] user2: I can't see figure for 90 or 100

The diagram is part of a scale drawing of a house. What is the length, in feet, of the side labeled ***x***?

You and your partner can each make ONE selection from the following list of hints. Use this information to provide your answer in the box below.

□ Value of A

□ Value of B



Current members: user1, user2

Type your message...

[17:54:51] user1: what is this horrible question asking us to do?!

[17:53:57] user2: cookies!

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

The diagram is part of a scale drawing of a house. What is the length, in feet, of the side labeled ***x***?

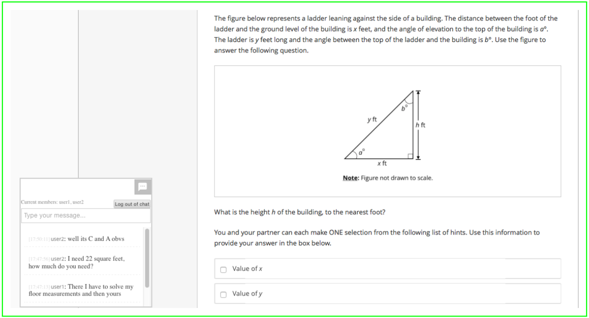
You and your partner can each make ONE selection from the following list of hints. Use this information to provide your answer in the box below.

□ Value of A

□ Value of B

□ Value of C

Hints / information requesting items



The figure below represents a ladder leaning against the side of a building. The distance between the foot of the ladder and the ground level of the building is *x* feet, and the angle of elevation to the top of the building is *a*°. The ladder is *y* feet long and the angle between the top of the ladder and the building is *b*°. Use the figure to answer the following question.

Note: Figure not drawn to scale.

What is the height *h* of the building, to the nearest foot?

You and your partner can each make ONE selection from the following list of hints. Use this information to provide your answer in the box below.

□ Value of *x*

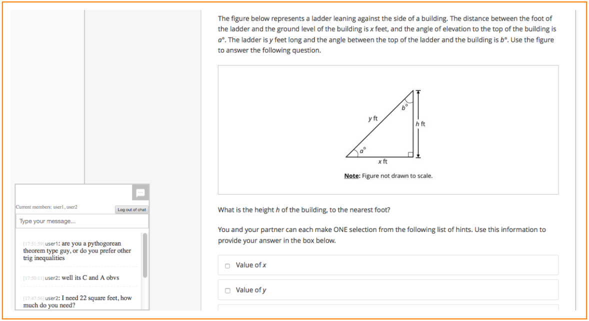
□ Value of *y*

Current members: user1 , user2

[17:50:11] user2: well its C and A obvs

[17:47:56] user2: I need 22 square feet, how much do you need?

[17:47:13] user1: There I have to solve my floor measurements and then yours



The figure below represents a ladder leaning against the side of a building. The distance between the foot of the ladder and the ground level of the building is *x* feet, and the angle of elevation to the top of the building is *a*°. The ladder is *y* feet long and the angle between the top of the ladder and the building is *b*°. Use the figure to answer the following question.

Note: Figure not drawn to scale.

What is the height *h* of the building, to the nearest foot?

You and your partner can each make ONE selection from the following list of hints. Use this information to provide your answer in the box below.

□ Value of *x*

□ Value of *y*

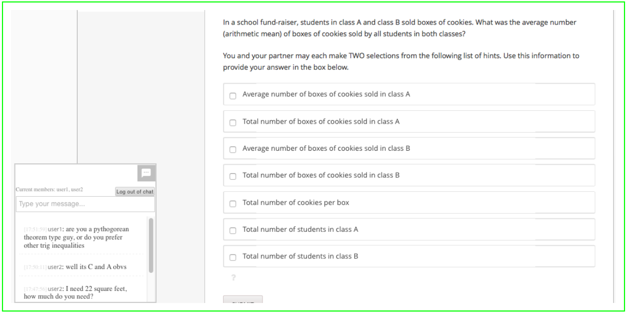
Current members: user1 , user2

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

[17:50:11] user2: well its C and A obvs

[17:47:56] user2: I need 22 square feet, how much do you need?

Hints / information requesting items



Current members: user1 , user2

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

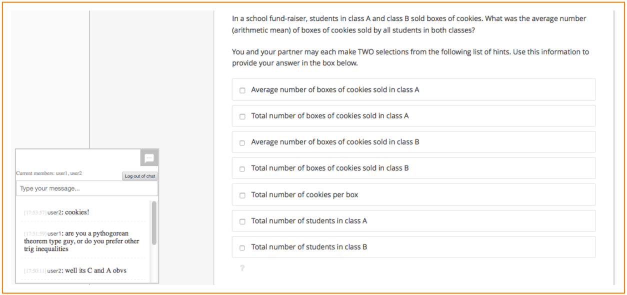
[17:50:11] user2: well its C and A obvs

[17:47:56] user2: I need 22 square feet, how much do you need?

In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

You and your partner may each make TWO selections from the following list of hints. Use this information to provide your answer in the box below.

- □ Average number of boxes of cookies sold in class A
- □ Total number of boxes of cookies sold in class A
- □ Average number of boxes of cookies sold in class B
- □ Total number of boxes of cookies sold in class B
- □ Total number of cookies per box
- □ Total number of students in class A
- □ Total number of students in class B



Current members: user1 , user2

[17:53:57] user2: cookies!

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

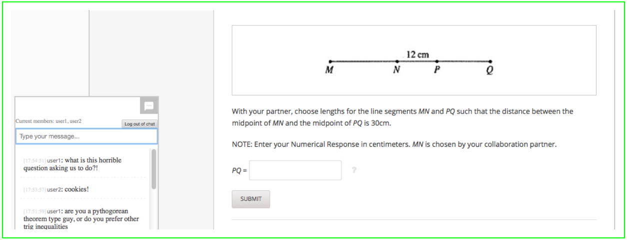
[17:50:11] user2: well its C and A obvs

In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

You and your partner may each make TWO selections from the following list of hints. Use this information to provide your answer in the box below.

- □ Average number of boxes of cookies sold in class A
- □ Total number of boxes of cookies sold in class A
- □ Average number of boxes of cookies sold in class B
- □ Total number of boxes of cookies sold in class B
- □ Total number of cookies per box
- □ Total number of students in class A
- □ Total number of students in class B

Multiple answer / negotiation items



Current members: user1 , user2

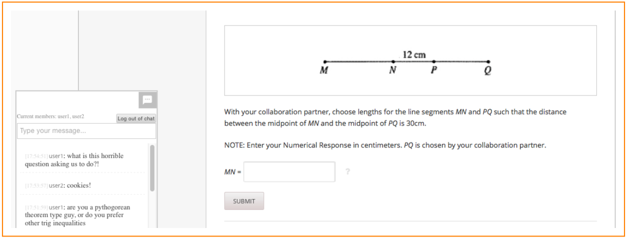
[17:54:51] user1: what is this horrible question asking us to do?!

[17:53:57] user2: cookies!

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

With your partner, choose lengths for the line segments *MN* and *PQ* such that the distance between the midpoint of *MN* and the midpoint of *PQ* is 30cm.

NOTE: Enter your Numerical Response in centimeters. *MN* is chosen by your collaboration partner.



Current members: user1 , user2

[17:54:51] user1: what is this horrible question asking us to do?!

[17:53:57] user2: cookies!

[17:51:59] user1: are you a pythogorean theorem type guy, or do you prefer other trig inequalities

With your collaboration partner, choose lengths for the line segments *MN* and *PQ* such that the distance between the midpoint of *MN* and the midpoint of *PQ* is 30cm.

NOTE: Enter your Numerical Response in centimeters. *PQ* is chosen by your collaboration partner.