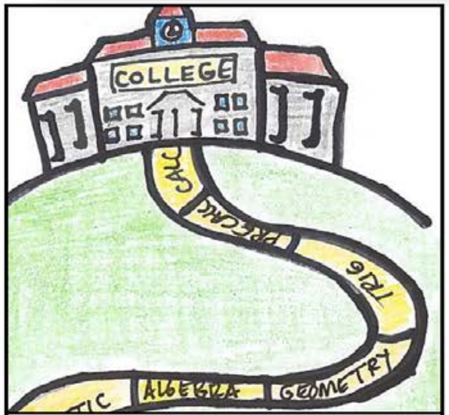
CALIBRATION OF CONFIDENCE JUDGMENTS IN ELEMENTARY MATHEMATICS:

MEASUREMENT, DEVELOPMENT, AND IMPROVEMENT

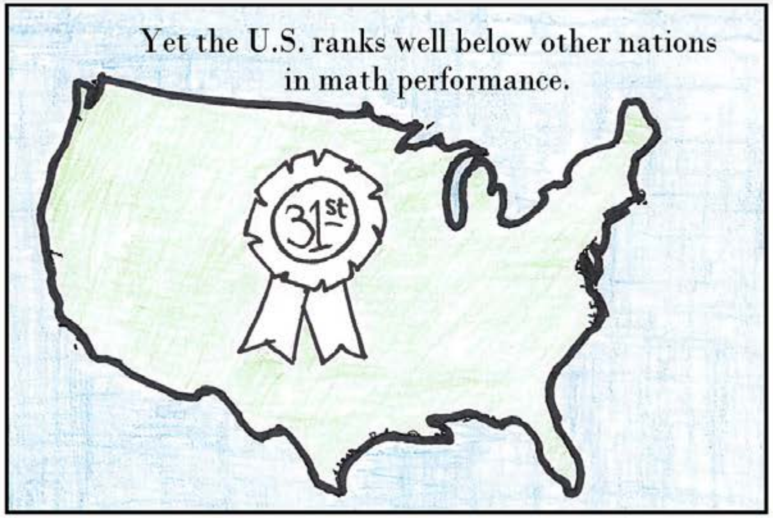
Teomara Rutherford

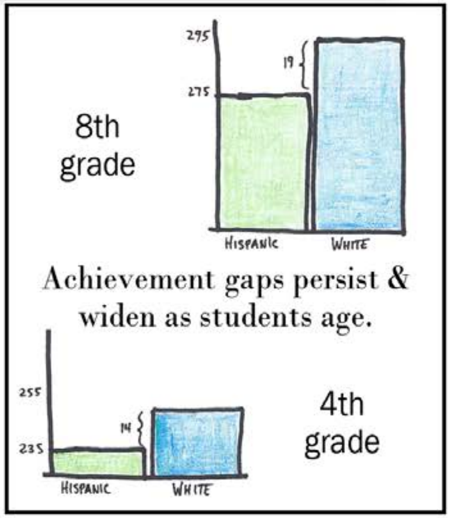
North Carolina State University

The Problem



Math is a pathway to higher education.


Yet the U.S. ranks well below other nations in math performance.



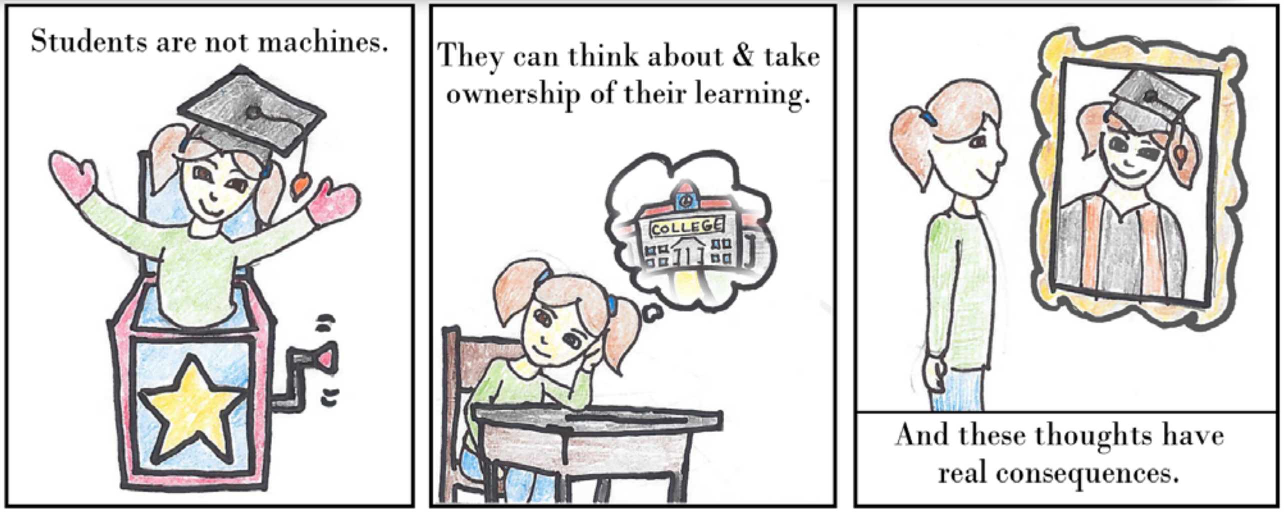
Achievement gaps persist & widen as students age.

How can we solve this problem?

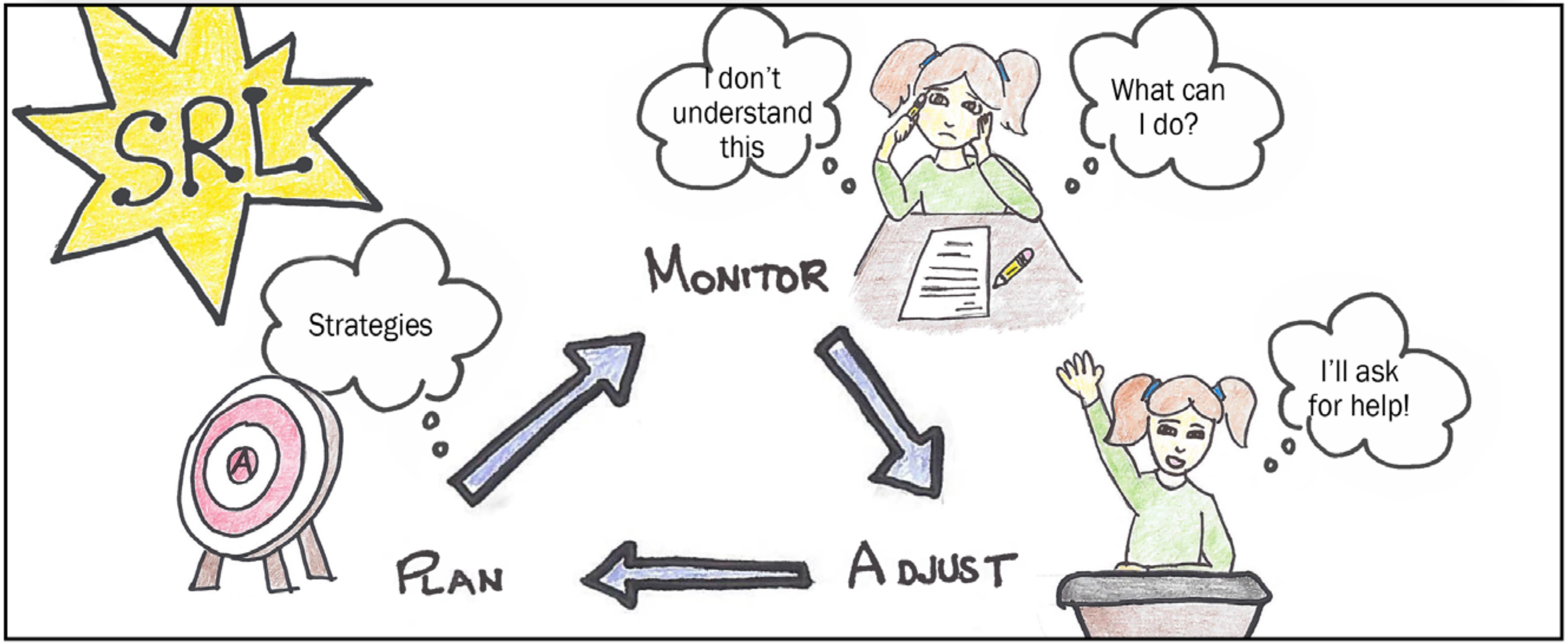
Teachers?

Schools?

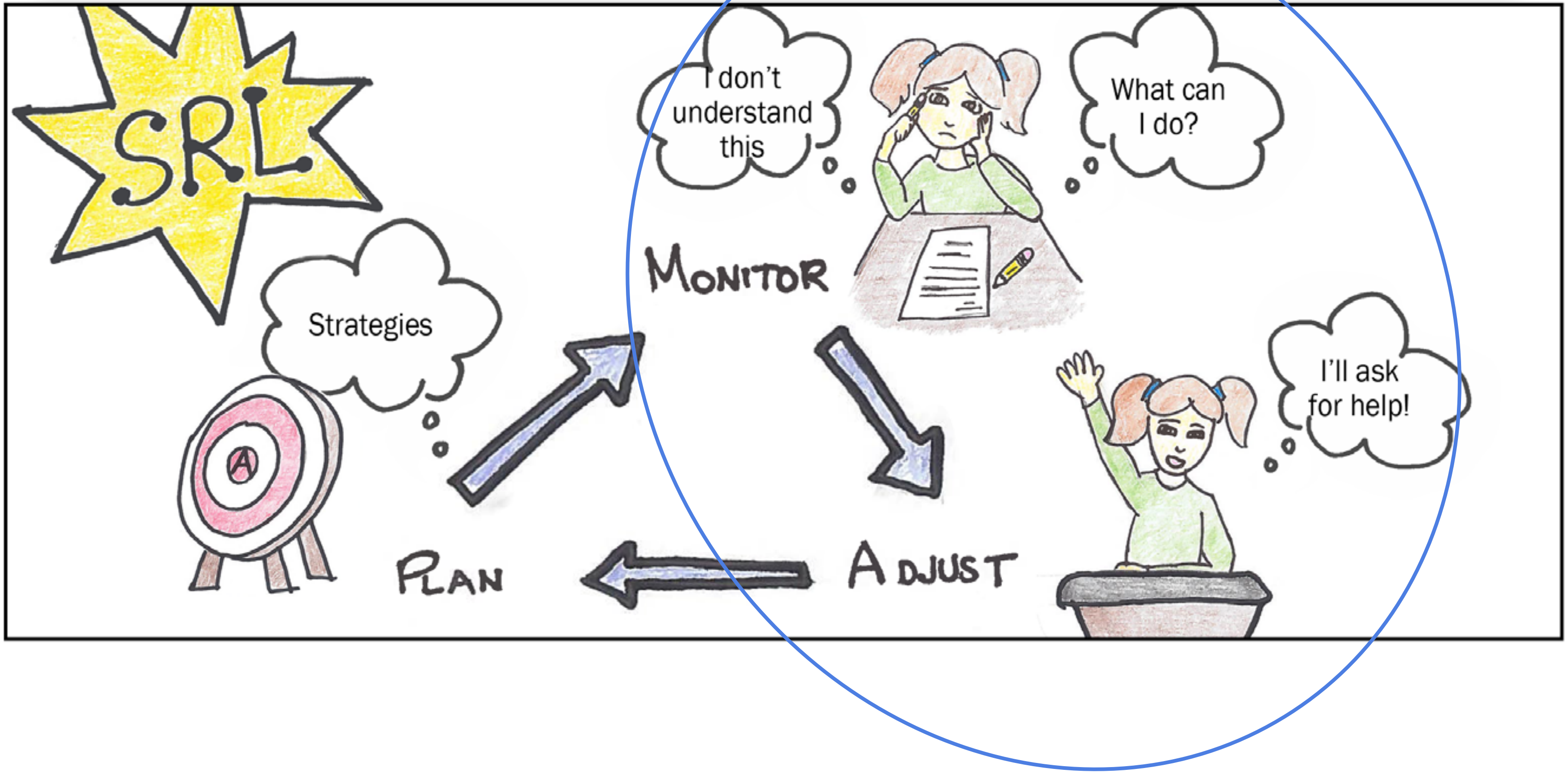
Curricula?


Students are not machines.
They can think about & take ownership of their learning.
COLLEGE
And these thoughts have real consequences.

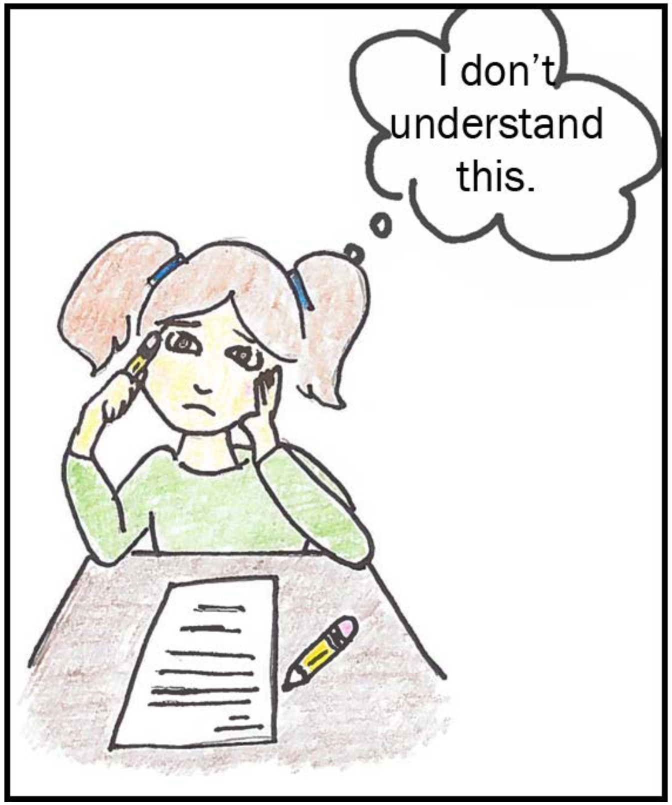
How do students regulate their learning?



How do students regulate their learning?

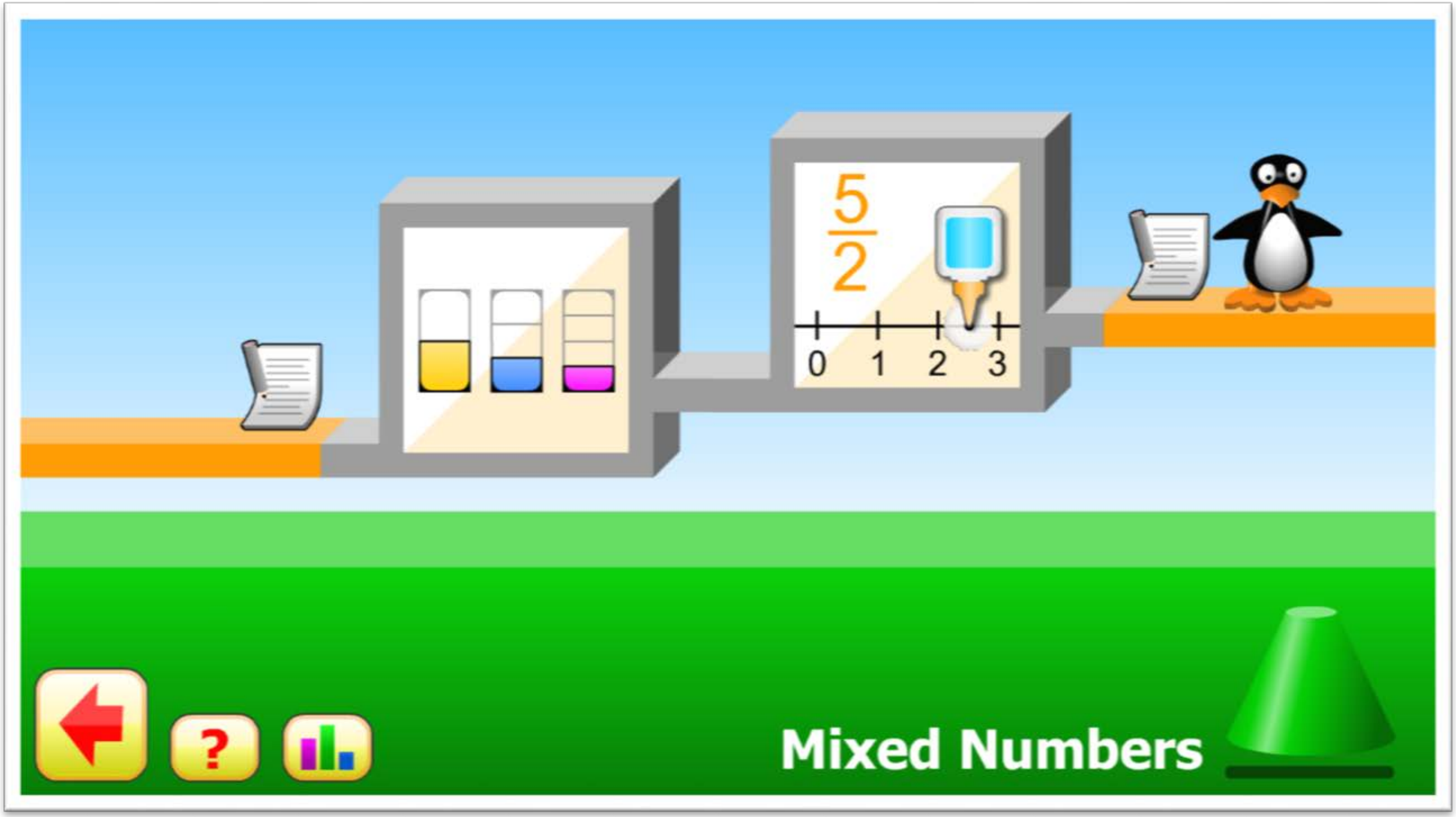


Calibration

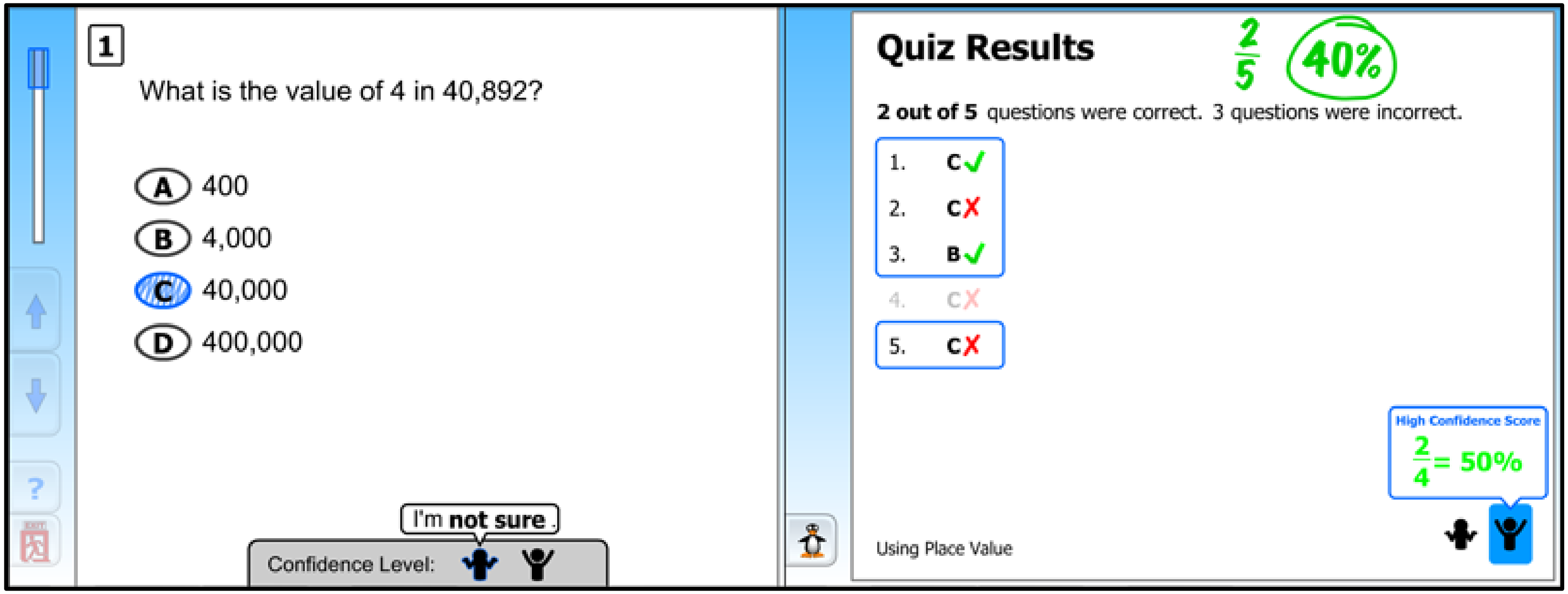


ST Math

ies INSTITUTE OF EDUCATION SCIENCES


5/2
0
1
2
3
?
Mixed Numbers

ST Math Quizzes



1

What is the value of 4 in 40,892?

- (A) 400
- (B) 4,000
- [✓ (C) 40,000]
- (D) 400,000

Quiz Results

$\frac{2}{5}$ 40%

2 out of 5 questions were correct. 3 questions were incorrect.

1.	C ✓
2.	C ✗
3.	B ✓
4.	C ✗
5.	C ✗

High Confidence Score

$\frac{2}{4}$ = 50%

Using Place Value

Does practice and feedback on calibration within ST Math improve student calibration accuracy?

Prior Work on Calibration

- More accurate calibration associated with higher achievement
- Content of material influences calibration accuracy
- Calibration can be improved through training, but this improvement often doesn’t translate to gains in achievement

Potential of Data

- Elementary students (previously understudied)
- Classroom activity
- Hierarchical domain of math
- Multiple measures of calibration and achievement for each student

Data Details

- ST Math
- Year-long curriculum, about 20 objectives per year
- 2nd through 5th grades
- 18 Southern California Schools
- > 4,000 students

How should I operationalize calibration?

A wrinkle from my committee

Research Questions

(1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments?

(2) Among these measures, which display the greatest predictive validity?

STUDY 1

	Correct	Incorrect
Confident	A Confident & Correct	B Confident & Incorrect
Not Confident	C Not Confident & Correct	D Not Confident & Incorrect

STUDY 1, QUESTION 1

Index	Formula
Sensitivity	$A/(A + C)$
Specificity	$D/(B + D)$
Simple Matching	$(A + D)/(A + B + C + D)$
G Index or Hamann coefficient	$(A + D) - (B + C)/(A + B + C + D)$
Odds Ratio	AD/BC
Goodman-Kruskal Gamma	$(AD - BD)/(AD + BC)$
Kappa	$2*(AD - BC)/[(A + B)(B + D) + (A + C)(C + D)]$
Phi	$(AD - BC)/[(A + B)(B + D)(A + C)(C + D)]^{1/2}$
Sokal Reverse	$[1 - [(A + D)/(A + B + C + D)]]^{1/2}$
Discrimination (d')	$z(A/(A + C)) - z(B/(B + D))$

Formulas as represented in Schraw et al., 2013.

	Correct	Incorrect
Confident	A Confident & Correct 62.5%	B Confident & Incorrect 12.5%
Not Confident	C Not Confident & Correct 12.5%	D Not Confident & Incorrect 12.5%

STUDY 1, QUESTION 1

	Correct	Incorrect
Confident	A Confident & Correct 62.5% (56%)	B Confident & Incorrect 12.5% (24%)
Not Confident	C Not Confident & Correct 12.5% (8%)	D Not Confident & Incorrect 12.5% (12%)

STUDY 1, QUESTION 1

Research Questions

(1) Which measures of calibration can accommodate real-world data of accuracy and confidence judgments?

(2) Among these measures, which display the greatest predictive validity?

Method

- Quizzes aggregated
- Posttest Accuracy = Calibration + Pretest Accuracy + Controls (demographics & game progress)
- Separate model for each of 10 measures
  - One model w/Sensitivity & Specificity together

STUDY 1, QUESTION 2

Results

(1)	(2)	(3)	(4)	(5)
Sensitivity	Specificity	Simple Match	G Index	Gamma
0.052***	-0.004	0.056***	0.056***	0.057***

(6)	(7)	(8)	(9)	(10)
Odds Ratio	Kappa	Phi	Sokal Reverse	Discrimination
0.021*	0.049***	0.054***	-0.052***	0.055***

(Combined)	
Sensitivity	Specificity
0.109***	0.074***

STUDY 1, QUESTION 2

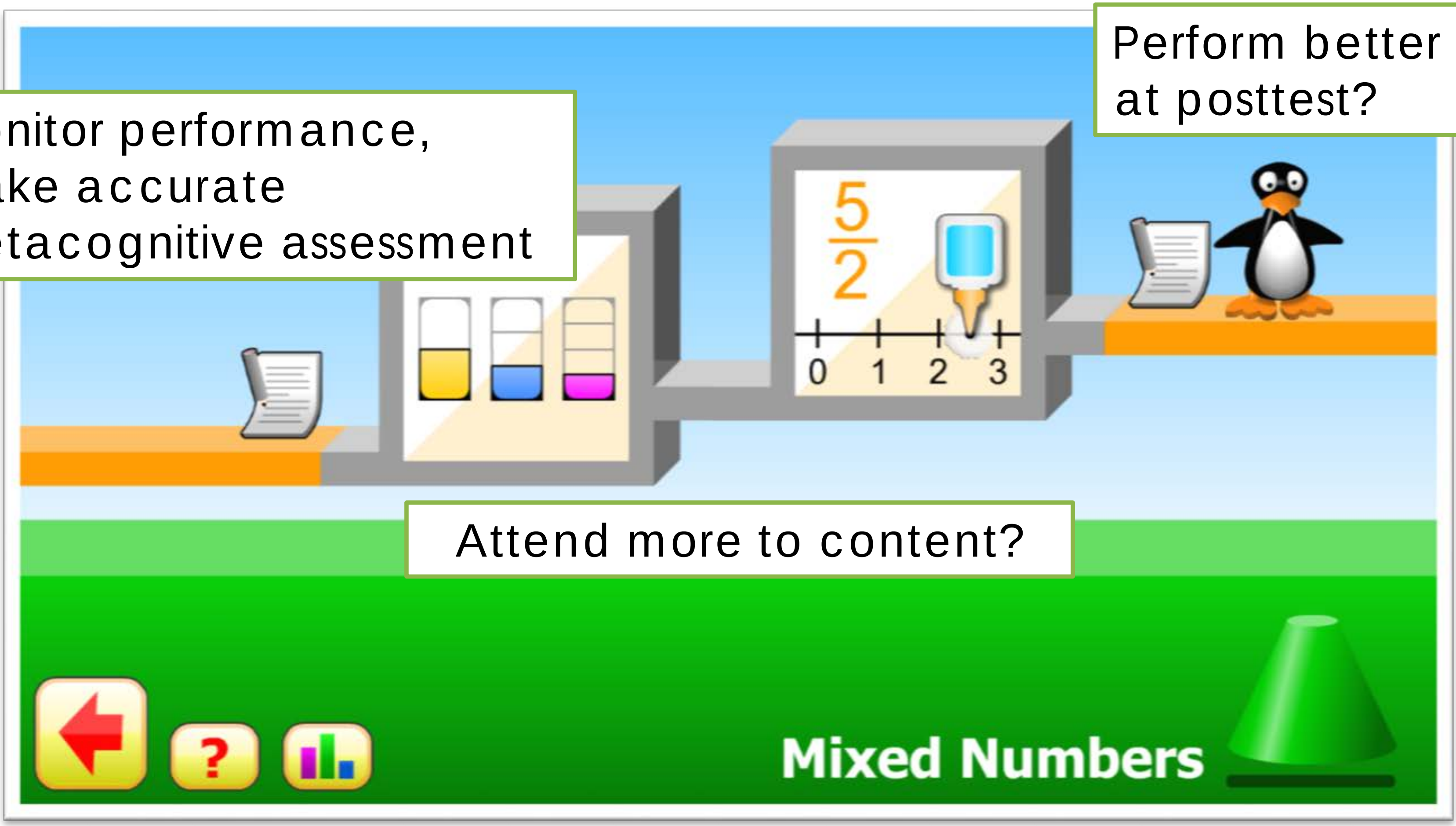
Conclusions

- Calibration researchers should consider problems of real data in choosing measures
- Sensitivity and Specificity should be considered—they are relatively robust to missing quadrants and when considered together, have strongest relations with achievement gain.

STUDY 1

WITHIN AND BETWEEN PERSON ASSOCIATIONS OF CALIBRATION AND ACHIEVEMENT

STUDY 2



Monitor performance, make accurate metacognitive assessment

Perform better at posttest?

Attend more to content?

STUDY 2

Research Question

Do students (within ST Math) make greater pre to posttest gains when better calibrated at pretest?

Method

- Calibration = Sensitivity & Specificity (accurate certainty and uncertainty)

- Random intercepts 2-level model
  - L1: Task x Person (quizzes)
  - L2: Person

- Student fixed effects (group-mean centering)

STUDY 2

Results

Level 1 (Objective)	
Sensitivity	Specificity
0.07***	0.02***

Level 2 (Student)	
Sensitivity	Specificity
0.09***	0.08***

Contextual Effect (Student Net Objective)	
Sensitivity	Specificity
0.02^{ns}	0.06***

STUDY 2

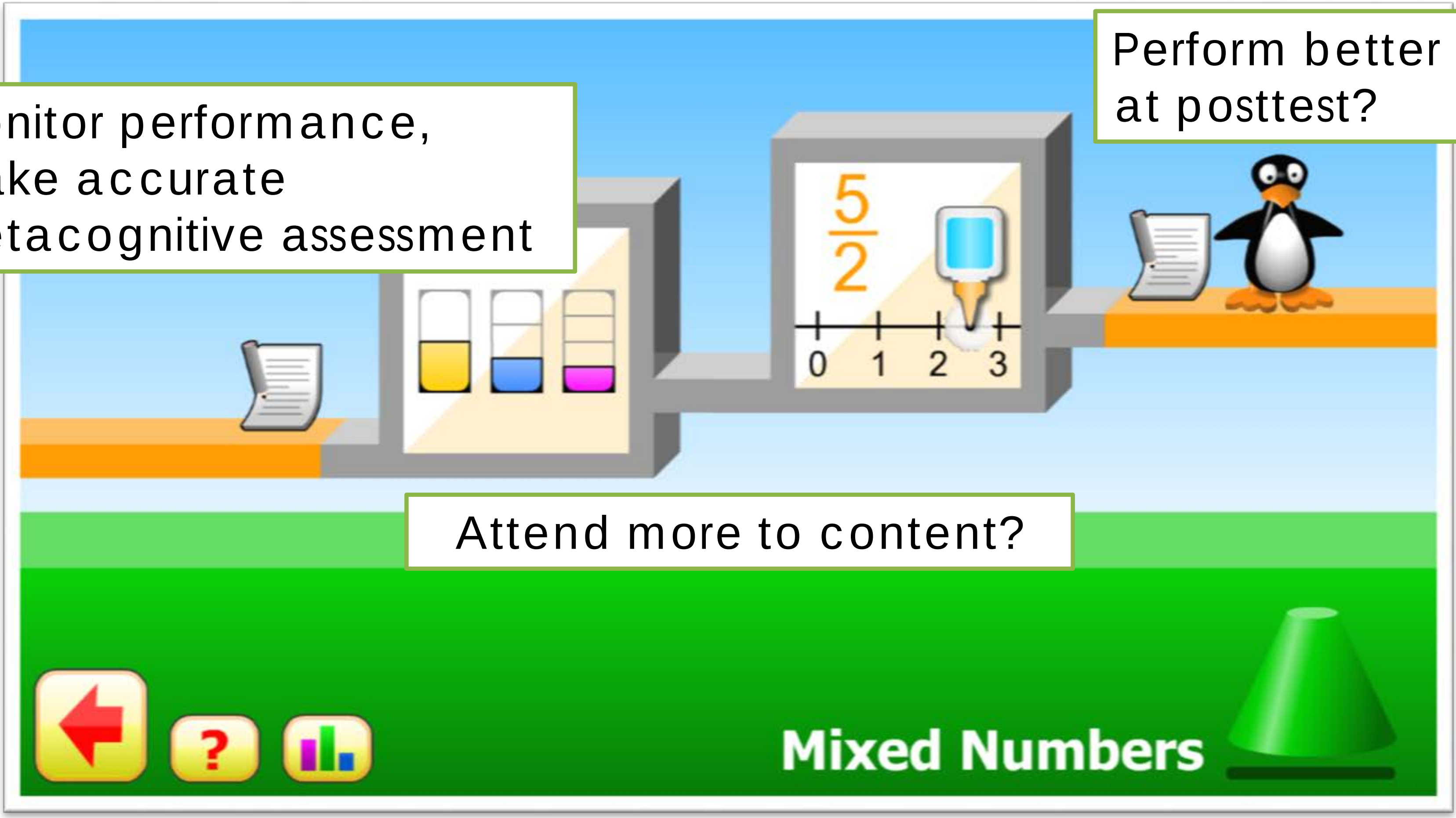
Replication

	Sensitivity	Specificity
Level 1	☑	☑
Level 2	☑	☑
Contextual	☑	☑

STUDY 2

Conclusions

- Small positive relation between calibration and performance both within and between students
- Sensitivity and Specificity had different associations with performance (at different levels)



Monitor performance, make accurate metacognitive assessment

Perform better at posttest?

Attend more to content?

Confident & Correct d=.10

Not Confident & Wrong d=.02

STUDY 2

CHANGES IN CALIBRATION:
IN RESPONSE TO INTERVENTION AND AS RELATED TO CHANGES IN ACHIEVEMENT

STUDY 3

Research Questions

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration?

(2) Is improvement in calibration accuracy linked to improvement in performance?

STUDY 3

Method

- Random variation in treatment start date
  - Early treatment group (ETG) started ST Math one year before Late treatment group (LTG)
- Posttest Calibration= Pretest Accuracy + Treatment Dummy + Controls
- Five commonly used measures of calibration

STUDY 3, QUESTION 1

	3	4	4
2008-2009	2009-2010	2010-2011	2011-2012
K	1st	2nd	3rd
1st	2nd	3rd	4th

STUDY 3, QUESTION 1

Results: ETG compared to LTG

	(1)	(2)	(3)	(4)	(5)
	Sensitivity	Specificity	Simple Match	Gamma	Discrimination
After Treatment (2011 to 2011)	⬇	⬆	⬇	⬇	⬇

STUDY 3, QUESTION 1

Results: ETG compared to LTG

	(1)	(2)	(3)	(4)	(5)
	Sensitivity	Specificity	Simple Match	Gamma	Discrimination
Before Treatment (2010 to 2011)	↓	*no sd*	↓	↓	↓
After Treatment (2011 to 2011)	↓	↑	↓	↓	↓

STUDY 3, QUESTION 1

Research Questions

(1) Can third and fourth grade students be trained to be more accurate in their calibration judgments through practice and feedback on accuracy and calibration?

(2) Is improvement in calibration accuracy linked to improvement in performance?

STUDY 3

Method

- Two types of analyses
  - Two related objectives (change scores)
  - Slopes of accuracy improvement on slopes of calibration improvement
- Within ST Math outcomes and state standardized test score outcomes
- Five calibration measures

STUDY 3, QUESTION 2

Results: ST Math

PAIRED QUIZZES

(1)	(2)	(3)	(4)	(5)
Sensitivity	Specificity	Simple Match	Gamma	Discrimination
0.07*	-0.07**	-0.04	0.0001	-0.005

SLOPES

(1)	(2)	(3)	(4)	(5)
Sensitivity	Specificity	Simple Match	Gamma	Discrimination
0.05	0.06	0.16	0.15	0.15

STUDY 3, QUESTION 2

Results: CSTs

PAIRED QUIZZES

(1)	(2)	(3)	(4)	(5)
Sensitivity	Specificity	Simple Match	Gamma	Discrimination
-0.05	0.04	0.01	-0.03	-0.01

SLOPES

(1)	(2)	(3)	(4)	(5)
Sensitivity	Specificity	Simple Match	Gamma	Discrimination
-0.001	0.01	0.03*	0.01	0.01

STUDY 3, QUESTION 2

Conclusions

- ST Math calibration practice may operate to increase uncertainty (Specificity)
- Change in calibration not associated with change in achievement in these data

STUDY 3

SUMMARY AND FUTURE DIRECTIONS

Key Findings

- Dual processes of calibration: certainty and uncertainty
- Calibration reflects elements of the Task x Person level and the Person level
- Calibration more complicated than represented in prior research

Future Directions

- Measurement
  - Dichotomous vs. more options
- Control
  - Student behaviors
- Aids to Malleability
  - Saliency of feedback
  - Direct instruction
- Experimental Manipulation
  - Separate out effect of ST Math and calibration feedback

Acknowledgements

My dissertation committee (& proposal committee):

George Farkas, Greg Duncan, Deborah Vandell, and Jacque Eccles; (Elizabeth Loftus, AnneMarie Conley)

Gregg Schraw and John Nietfeld for feedback

MIND Research Institute, Orange County Department of Education, and the students and teachers within the study

Funders: IES (Grant R305A090527) and NSF GRFP (Grant DGE-0808392).

Questions?

Teya Rutherford
taruther@ncsu.edu