CONNECTING EVALUATION AND COMPUTING EDUCATION RESEARCH: WHY IS IT SO IMPORTANT?

R·I·T Adrienne Decker

KNOX COLLEGE Monica M. McGill

Jason Ravitz Google

Eric Snow SRI International

Rebecca Zarch SAGEFOX CONSULTING GROUP

Session Outline

- Introduction
- Jason Ravitz – Evaluation Wrecking Crew → CS Impact Network
- Eric Snow – Valid Measures Matter
- Rebecca Zarch – SageFox Consulting Group Projects
- Q & A

Introduction (Why are we here?)

- Our (Decker & McGill) current project has a lot to do with evaluation
  - *But we are not evaluators*

- Wanted to bring together evaluators to discuss importance of evaluation in our CS Ed research community
  - *Share their experiences and insight*
  - *Inspire us to work harder at better evaluation*

EVAL WRECKING CREW → CS IMPACT NETWORK

Jason Ravitz

Google

Outline

- What is evaluation?
- Relationship to research
- Examples from Google
- The Eval Wrecking Crew → CS Impact Network

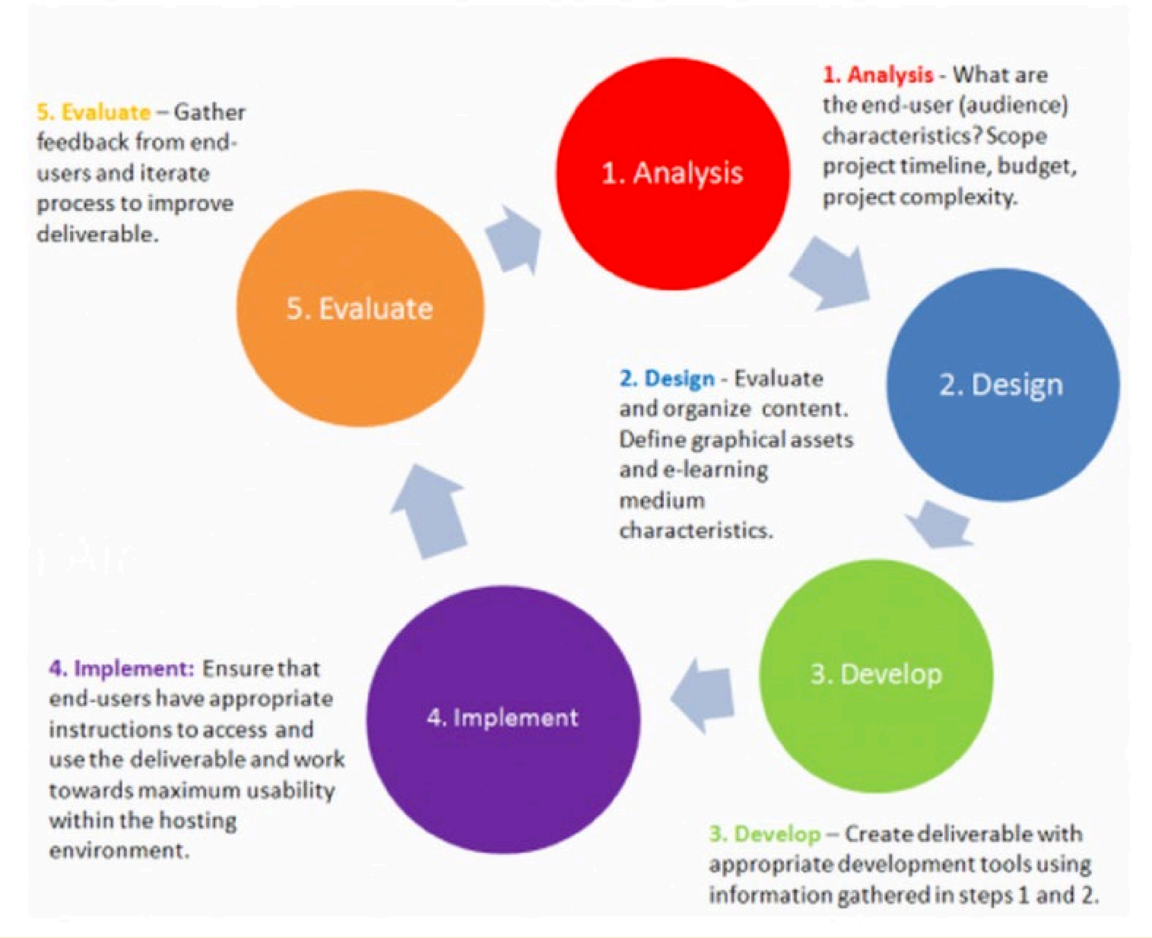
What is evaluation?

- **Systematic collection of information** to
  - *make judgments*
  - *improve programs*
  - *inform program decision making, and*
  - *increase understanding*
    - (Michael Patton, 2008)

- **Determining merit, worth, value** or significance for stakeholders
  - *(American Evaluation Association)*

Evaluation is part of good design..

Instructional Design



1. Analysis - What are the end-user (audience) characteristics? Scope project timeline, budget, project complexity.

2. Design - Evaluate and organize content. Define graphical assets and e-learning medium characteristics.

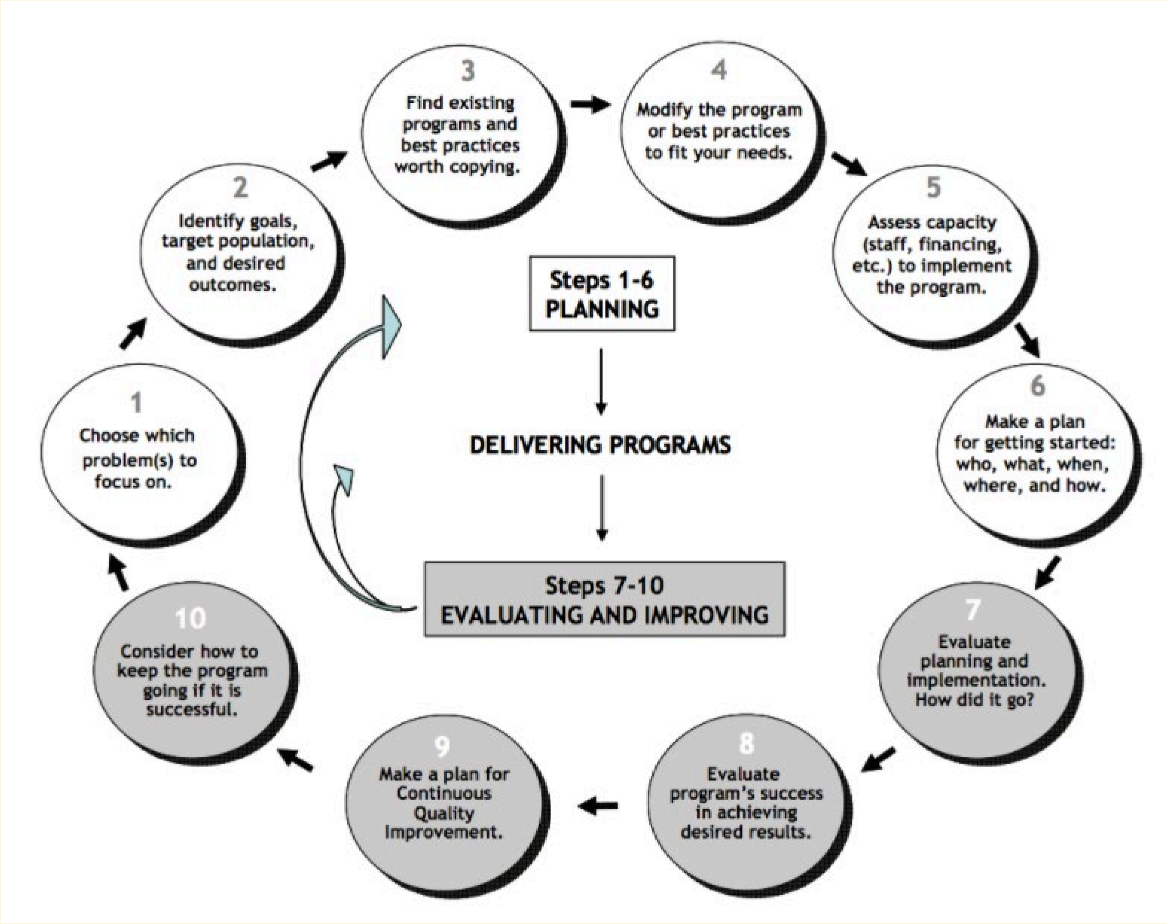
3. Develop – Create deliverable with appropriate development tools using information gathered in steps 1 and 2.

4. Implement: Ensure that end-users have appropriate instructions to access and use the deliverable and work towards maximum usability within the hosting environment.

5. Evaluate – Gather feedback from end-users and iterate process to improve deliverable.

Morrison, G., Ross, S., Kemp, J. (2004). *Designing effective instruction.*

Program Design



1 Choose which problem(s) to focus on.

2 Identify goals, target population, and desired outcomes.

3 Find existing programs and best practices worth copying.

4 Modify the program or best practices to fit your needs.

5 Assess capacity (staff, financing, etc.) to implement the program.

6 Make a plan for getting started: who, what, when, where, and how.

7 Evaluate planning and implementation. How did it go?

8 Evaluate program's success in achieving desired results.

9 Make a plan for Continuous Quality Improvement.

10 Consider how to keep the program going if it is successful.

Rand (2007). Getting To Outcomes™ 10 steps for Achieving Results-Based Accountability. p. 2-3

Reasons to Evaluate

- Improve
  - *Can we do better with design/implementation?*
- Learn
  - *What is/isn’t effective?*
  - *What are key success factors?*
  - *What are diverse perspectives?*
- Judge
  - *Are we accountable?*
  - *Should we change?*
  - *What should we invest in more?*

Evaluation = A form of research

- Tied to program development
- Rooted in organizational contexts
- **Designed** to inform decisions
- Used for accountability, to judge merit or worth
- Focused on diverse stakeholders
- **Yielding lessons for improving ⟵ the most "research-y"**
  - e.g., what works, for whom, under what conditions

Common Practices for R&E

- Developing and studying interventions (w/educators, e.g., RPPforCS)
- Establishing baseline measures
- Conducting Literature Reviews
  - *to identify questions, methods, measures, sources of error, etc.*
- Developing, validating and re-using measures
- Analyzing data
- Developing narratives / storytelling
- Reporting results
- Struggling with causality, equity, dissimilar conditions, etc.

Research supports evaluation by…

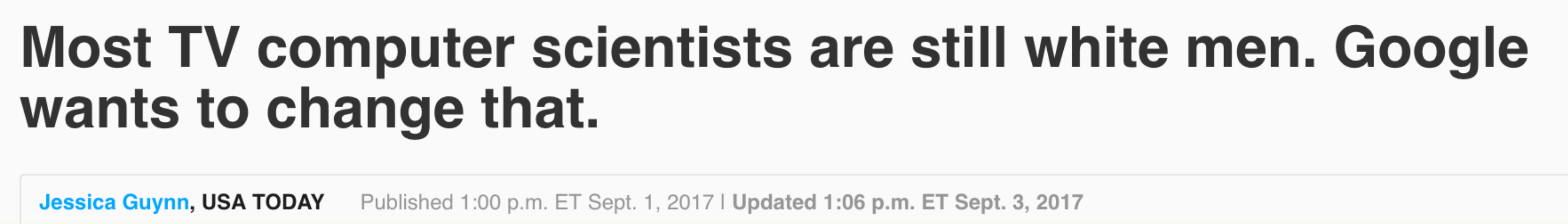
- **Building** theories to support cumulative learning and change

- **Addressing basic and (sometimes) practical questions**
  - e.g., what influences interests in CS?

- **Developing methods/measures** to increase confidence, reduce error, and increase accuracy (e.g., sampling, open-ended questions, peer review)

Examples: CS in Media

- **Research:** Exposure to CS is important, but so are self-perceptions, career-perceptions and social encouragement.
  - Predicts 60% of interest in CS
    - "Women who Choose" study → g.co/cseduresearch
- **Evaluation:** CS in Media program impacts
  - Hyperlinked (used same items)
    - *"Girls who have seen the first season are 11% more likely to be interested in computer science"*
      - USA Today article → tinyurl.com/csim-usatoday

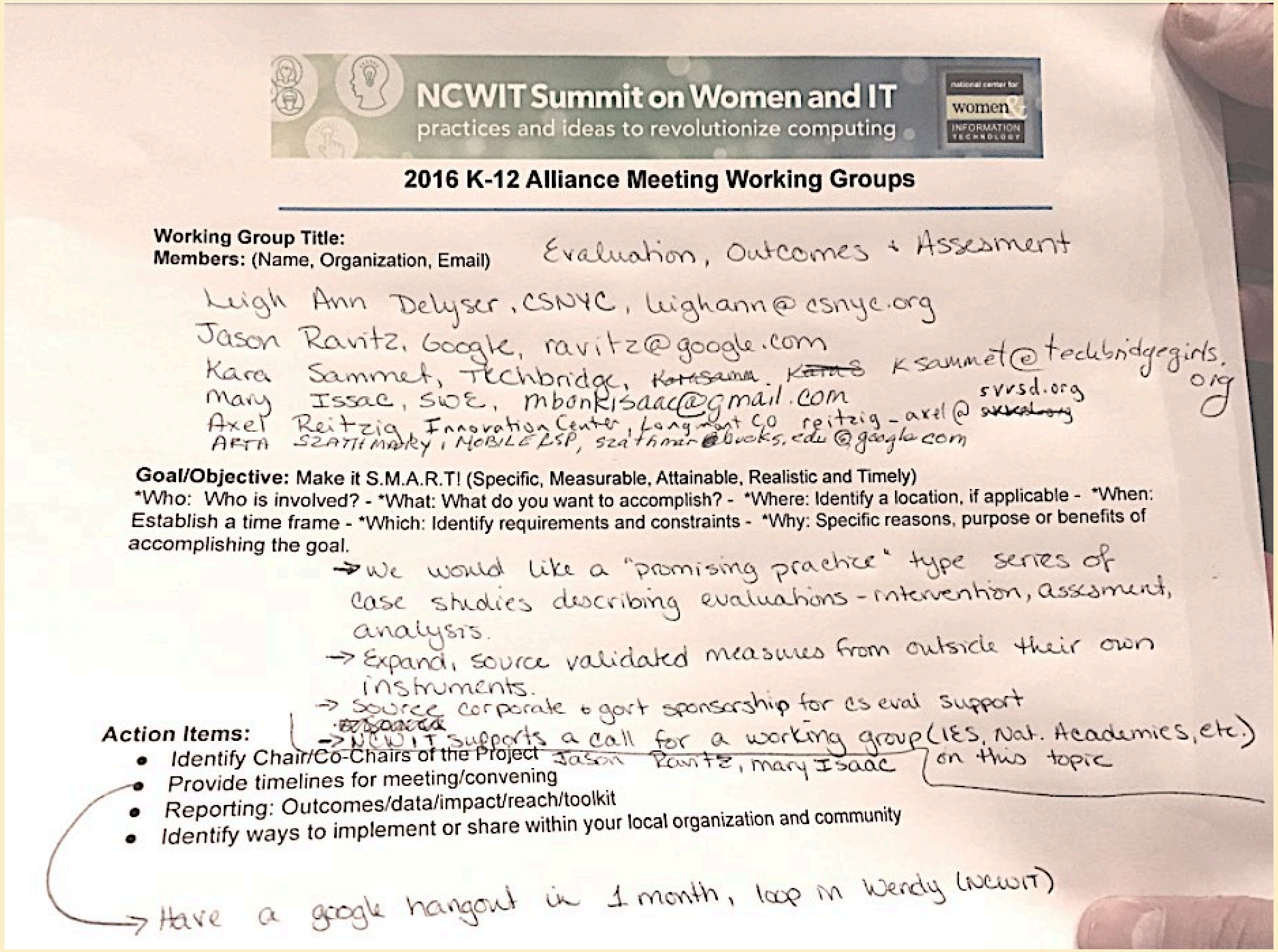


Most TV computer scientists are still white men. Google wants to change that.

Jessica Guynn, USA TODAY   Published 1:00 p.m. ET Sept. 1, 2017 | Updated 1:06 p.m. ET Sept. 3, 2017

Eval Wrecking Crew → CS Impact Network

Humble beginnings (N=6)



NCWIT Summit on Women and IT
practices and ideas to revolutionize computing

2016 K-12 Alliance Meeting Working Groups

Working Group Title: Evaluation, Outcomes + Assesment
Members: (Name, Organization, Email)

Leigh Ann Delyser, CSNYC, leighann@csnyc.org
Jason Ravitz, Google, ravitz@google.com
Kara Sammet, Techbridge, ~~Karasamm~~, ~~Karas~~ ksammet@techbridgegirls.org
Mary Issac, SWE, mbonkisaac@gmail.com
Axel Reitzig, Innovation Center, Longmont CO reitzig_axel@ ~~[illegible]~~ svvsd.org
ARTA SZATHMARY, MOBILE CSP, szathmar@books.edu @google.com

Goal/Objective: Make it S.M.A.R.T! (Specific, Measurable, Attainable, Realistic and Timely)
*Who: Who is involved? - *What: What do you want to accomplish? - *Where: Identify a location, if applicable - *When: Establish a time frame - *Which: Identify requirements and constraints - *Why: Specific reasons, purpose or benefits of accomplishing the goal.

→ We would like a "promising practice" type series of case studies describing evaluations - intervention, assesment, analysis.

→ Expand, source validated measures from outside their own instruments.

→ Source corporate + govt sponsorship for cs eval support

~~resources~~

→ NCWIT supports a call for a working group (IES, Nat. Academies, etc.) on this topic

Action Items:

- Identify Chair/Co-Chairs of the Project Jason Ravitz, Mary Isaac
- Provide timelines for meeting/convening
- Reporting: Outcomes/data/impact/reach/toolkit
- Identify ways to implement or share within your local organization and community

→ Have a google hangout in 1 month, loop in Wendy (NCWIT)

A growing number of efforts

- **Groups we started working with…**
  - NCWIT K-12 Alliance
  - AEA STEM TIG
  - NSF Grants
    - CSONIC (csonic.org)
    - Pre-College Computing (csedresearch.org)
    - STEM Evaluation Community
  - National Girls Collaborative Project (CS OPEN)

- Evaluation Wrecking Crew was formed **to work together** on common problems (measurement, capacity, design, etc.)

Strong Growth

- Our volunteer-led initiative was productive and attracted many participants.
  - ***5*** *members to start*
  - ***14*** *members EOY 2016,*
  - ***34*** *members in 2017,*
  - ***53*** *members, including from* ***10*** *universities in 2018.*

Oak Ridge Associated Universities (a ***121****-member university consortium) and the American Evaluation Association have also made a commitment to help develop our repository.*

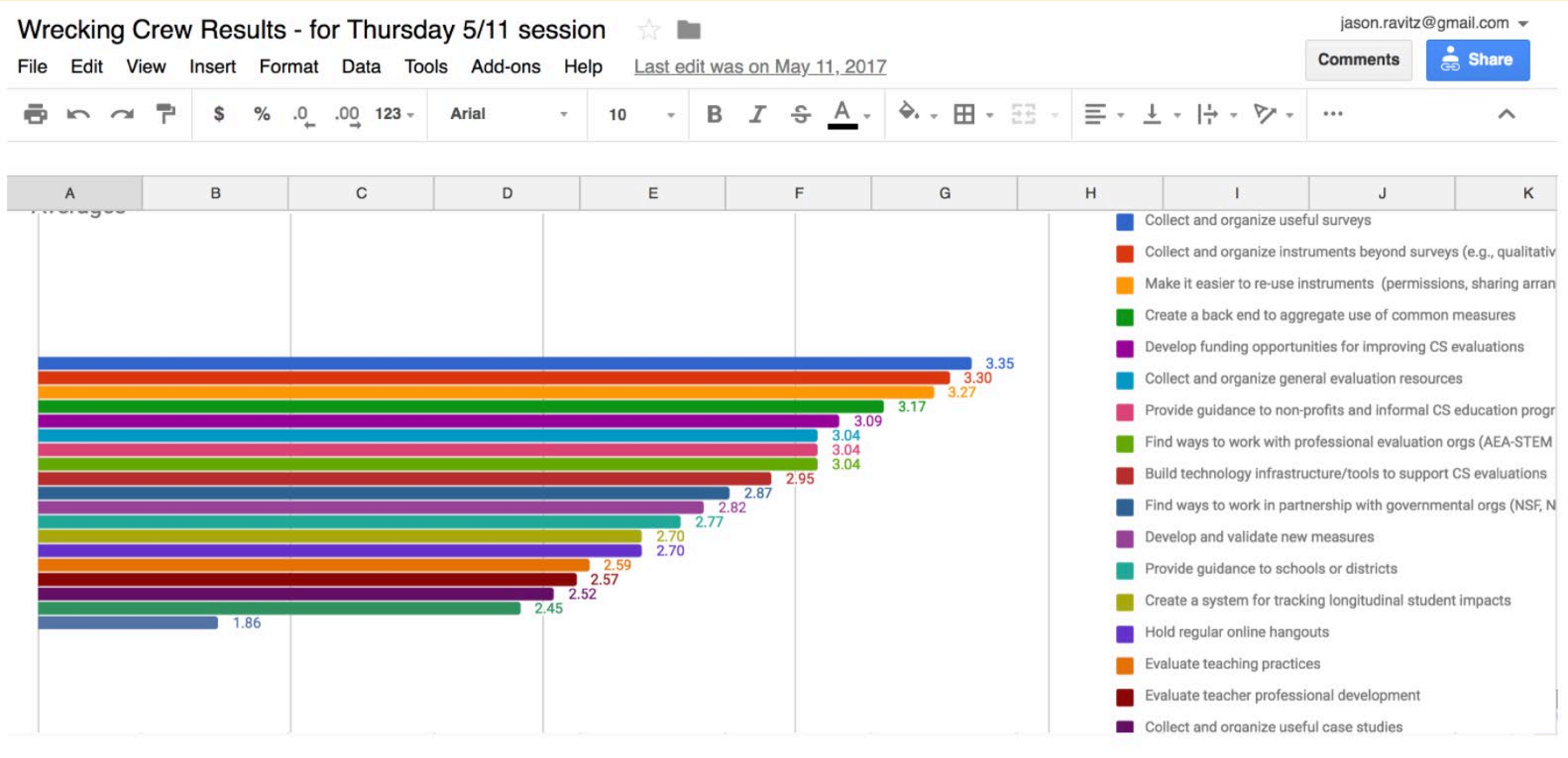
Supporting Mechanisms

Convening	Zoom remote meetings (after we maxed out Hangouts)
Tools	Empowerment Exercise** Feedback forms** Worksheets**
Monitoring (to track and measure progress over time)	Evaluation Dashboards with Goals, milestones, baselines, and actual performance

Accomplishments to date

- Bi-monthly meetings (average 10+)
- A needs assessment survey and analysis
- A meta-repository review form and analysis
- Evaluation repository (with growing CS emphasis)
- An empowerment exercise used for multiple programs
  - *Eval Wrecking Crew (NCWIT, etc.)* *tinyurl.com/wcrew-ee*
  - *CSONIC workshop for NSF evaluators* *tinyurl.com/cise-eval*
  - *AERA workshop for education leaders* *tinyurl.com/eetemplate*

Needs assessment results



Empowerment Process

- Step 1: **MISSION**
  - *Identify mission statements*
- Step 2: **TAKING STOCK** –
  - *BRAINSTORM - List of important things the group is doing*
  - *PRIORITIZE LIST - Vote 5 times for the most important things we should assess as a group*
  - *RATE & DISCUSS - How well are we are doing using a 1 (low) to 10 (high) scale? Then discuss why key ratings were made for a few activities.*
- Step 3: **PLAN FOR THE FUTURE** –
  - *Specify goals, strategy, and credible evidence (basis for rating in Taking Stock can be used to inform strategies in Plans for the Future)*
- Step 4: **MONITOR** –
  - *Evaluation dashboards, including goals, strategies, and evidence*

Initiatives (resulting from EE exercise)

Goal (purpose)	**Activities** (to accomplish goals)
Creating a centralized hub	Web Page Design list and link members
Provide STEM evaluation resources repository	Wrecking Crew survey; CSONIC needs assessment Repository: Design, test, refine + PEAR resources @ Harvard Invite dialogue and critique
Educating policy and decision makers	Dissemination: presentations, articles, chapters, blogs, workshop, AEA, AERA
Teaching about effective measures	Link and provide training resources, including web sites (e.g. (Better Evaluation)
Inviting corporate stakeholders	Online and in person meetings Sharing agendas and priorities Making value explicit

New Vision: CS Impact Network

Updated mission is to…

- Build evaluators’ capacity
- Improve the quality of computer science education to help…
  - *Students actualize their potential,*
  - *Teachers deliver quality programs,*
  - *Administrators support teaching and learning,*
- Produce a digitally prepared, technologically literate, and productive workforce.

Sustainability (No longer Google-led)

Seeking funding (internal or external) for

- Coordination and Administration of Consortium.
- Facilitation of Capacity Building Exercises.
- Hub
- Repository
- Educating CS Community
- Corporate, Foundation, and Philanthropic Stakeholders
- Holding a Summit

Example Resources

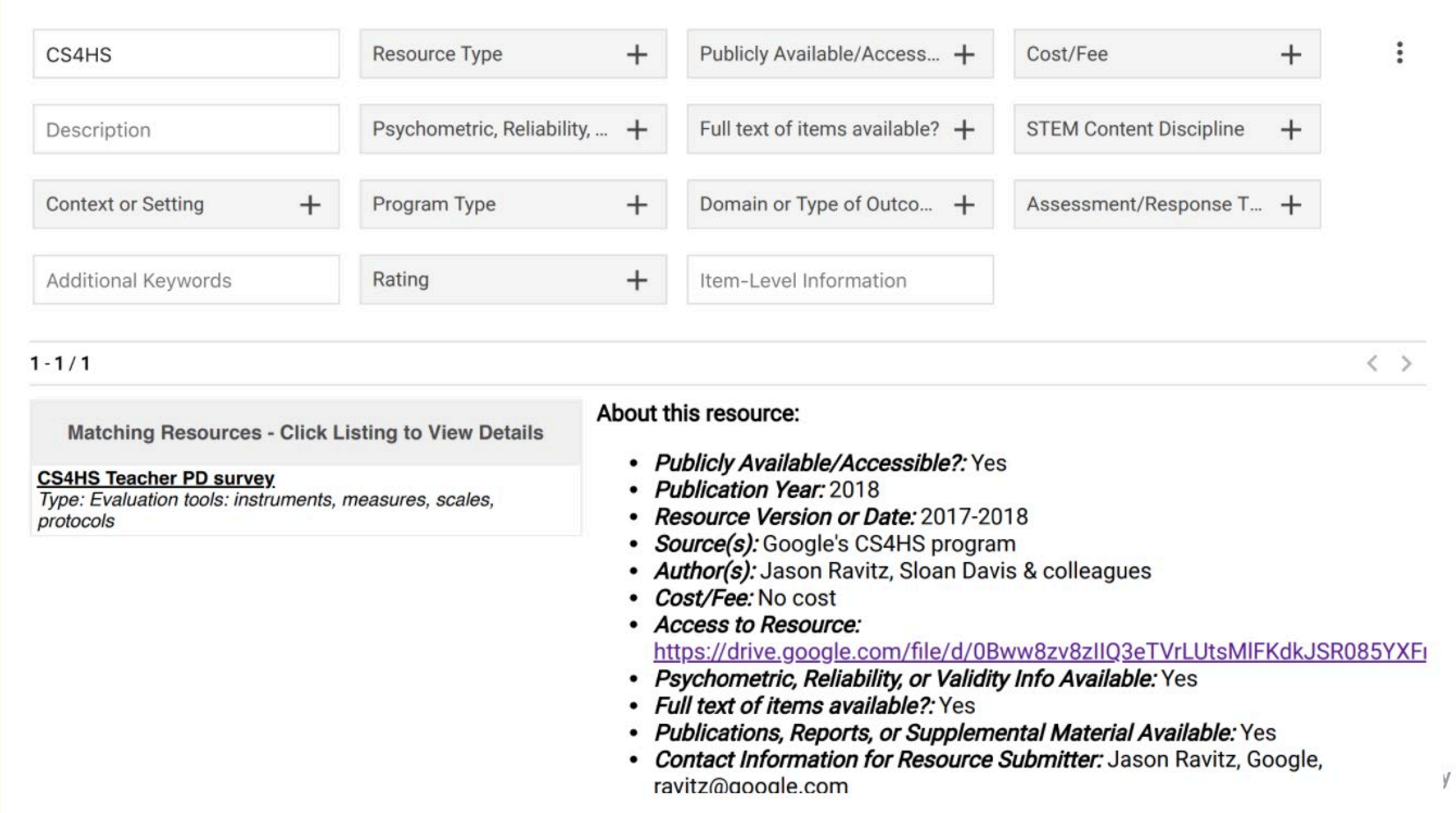
- **Evaluation Planning Worksheet**
  - tinyurl.com/evalworksheet-google

- **Edu on Air (Empowering leaders with evaluation best practices)**
  - tinyurl.com/ravitz-eduonair

- **Empowerment Evaluation Exercise**
  - tinyurl.com/eeblank

- **21st Century Teaching Survey**
  - academia.edu/5901608

Pilot for Repository (in Awesome Tables)



CS4HS

1 - 1 / 1

Matching Resources - Click Listing to View Details

CS4HS Teacher PD survey
Type: Evaluation tools: instruments, measures, scales, protocols

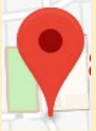
About this resource:

- ***Publicly Available/Accessible?:*** Yes
- ***Publication Year:*** 2018
- ***Resource Version or Date:*** 2017-2018
- ***Source(s):*** Google's CS4HS program
- ***Author(s):*** Jason Ravitz, Sloan Davis & colleagues
- ***Cost/Fee:*** No cost
- ***Access to Resource:*** https://drive.google.com/file/d/0Bww8zv8zIIQ3eTVrLUtsMlFKdkJSR085YXFı
- ***Psychometric, Reliability, or Validity Info Available:*** Yes
- ***Full text of items available?:*** Yes
- ***Publications, Reports, or Supplemental Material Available:*** Yes
- ***Contact Information for Resource Submitter:*** Jason Ravitz, Google, ravitz@google.com

Presentations

- American Evaluation Association

  1. An Evolving Repository of STEM Evaluation Resources
  2. Building a CS/STEM Evaluation Learning Community
  3. Building Evaluative Capacity of Out-of-School Organizations
  4. The National Girls Collaborative Project and Google (2016)



- SIGCSE

  1. Connecting Evaluation and Computing Education Research
  2. Repositories you shouldn't be living without (tomorrow @)

- AERA (April, 2018 - accepted)

  1. Building a Virtual CS/STEM Evaluation Learning Community

VALID MEASURES MATTER

Eric Snow

Senior Education Researcher

SRI International

Significance & Need

Computer science is spreading throughout the US K-12 system

Increasing demand for assessments that support valid inferences about student learning

Development of high-quality assessments has not kept pace with the spread of CS programs/curricula throughout the US K-12 system

Significance & Need

- **Teachers in the introductory CS courses** >>> how should I adapt instruction to meet my students' needs?
- **Teachers in advanced CS courses** >>> how well are students prepared for advanced work and where do they need extra help?
- **Principals** >>> is my school offering rigorous CS courses?
- **Stakeholders** >>> what CS knowledge and skills students are developing?

Assessment Challenges

Challenge #1: Understanding the Domain

>>> What is important for computer scientists to know and be able to do? What are the important **practices** of CS?

Challenge #2: Developing Authentic Representations

>>> How can we develop tasks/situations that elicit evidence of computational thinking
practices?

Challenge #3: Eliciting Valid Evidence

>>> Does the evidence support the inferences we want to make about computational thinking practices?

Assessment Challenges

Challenge #3: Eliciting Valid Evidence

>>> To what extent does ***evidence support the inferences*** we
want to make about computational thinking practices?

Test Validity

- Construct validity
- Criterion-related validity
- Convergent and discriminant validity
- Predictive validity
- …

Test Validity

- Construct validity
- Criterion-related validity
- Convergent and discriminant validity
- Predictive validity
-

NOPE

Test Validity

The latest thinking in test validity focuses on **supporting assessment inferences** through collecting and integrating **different types of evidence**:

- ➢ Test Content
- ➢ Internal Structure
- ➢ Response Processes
- ➢ Relations to other Variables
- ➢ Test Use

Test Validity Evidence

Type	Key Evidence
Test Content	Degree of alignment between test questions and learning objectives, standards and other guiding design documents >>> Expert Review
Internal Structure	Extent to which test scores support theoretical structure of assessment >>> Reliability, Factor Analysis, Item Characteristics
Relationship with other Variables	Extent to which test scores are related to other variables >>> Correlations

Test Validity Evidence

Type	Key Evidence
Response Processes	Extent to which student cognitive processes while completing test questions align with question design expectations >>> Think-Aloud Interviews / Cognitive Labs
Test Use (Consequences)	Extent to which consequences of the use of the score are congruent with the proposed uses of the assessment. >>> Predictive correlational analysis, qualitative investigations

Context – *Exploring Computer Science* (ECS)

- Pre-AP, introductory CS course
- Late middle school / early high school
- Six, six-week units
- Focus on **equity**
- A central tenet of ECS pedagogy is **inquiry-based learning:** core concepts learned through induction, teaching through guided inquiry, and open-ended assessments

Designing & Developing Assessments for Exploring Computer Science

Snow, E., Tate, C., Rutstein, D., Bienkowski, M. (2017). *Assessment design patterns for computational thinking practices in Exploring Computer Science.* (SRI technical report). Menlo Park, CA: SRI International.

Bienkowski, M., Snow, E., Rutstein, D. W., & Grover, S. (2015). *Assessment design patterns for computational thinking practices in secondary computer science: A first look*. (SRI technical report). Menlo Park, CA: SRI International.

Available: https://pact.sri.com/resources.html

Piloting & Validating Assessments for Exploring Computer Science

- Pilot 1 2014-2015, Pilot 2 2015-2016
- ECS teachers from across the U.S. including Los Angeles, Chicago, and New York
- Early on collected validity evidence based on test content and student responses processes to help us refine and improve the assessments
  - *Test content >>> expert review of alignment between the knowledge and skills, the curriculum learning goals, and CT practices*
  - *Student response processes >>> cognitive think-aloud interviews with students participating in the pilot testing activities*

Scoring & Inter-Rater Reliability

- Researchers were trained on the rubrics
- Each assessment was scored by two different scorers with a third scorer scoring if there were discrepancies in the scores
- Inter-rater reliability was high, with over 90% agreement between raters for most of the tasks
- Tasks for which the reliability was lower were revised either by modifying the item to clarify what was expected or by modifying the rubric

Descriptive Statistics

- ~ 40% female/ 60% males, ~50% Hispanic/Latino (49.28%)
- Average total scores in the 60 - 70% range across the assessments
- Female and male students had comparable average scores on the assessments
- Score distributions were slightly negatively skewed, indicating more students scored at the high end of the score distributions.

Validity Evidence Based on Internal Structure - Inferences

- Inference #1: The unit assessments measure one main construct (unidimensionality)
- Inference #2: The assessment questions are of appropriate difficulty for students
- Inference #3: The assessments can help teachers distinguish students at different ability levels

Validity Evidence Based on Internal Structure - Evidence

- Moderate to high levels of reliability (.66 - .84)
- Factor analysis supported expected structure of unit and cumulative assessments
- Moderate task difficulty levels, with the index ranging from .58 to .67
- High discriminating power for tasks/items with medium levels of difficulty

Discussion

Validity evidence based on internal structure is particularly promising:

- *tasks within each unit assessment are all measuring one general construct*
- *assessments best suited for differentiating students of average ability*

Discussion

Next Steps

- Examine whether validity results hold w/ larger sample and schools from different contexts
- Developing additional assessment tasks, particularly those with easy and hard levels of difficulty to improve utility across wider range of ability levels
- Item Response Theory (IRT) and Testlet Response Theory (TRT) analyses

Conclusions

Important effort to apply principled assessment design methods and contemporary test-validity standards to guide the development, piloting and validation of assessments of CTPs

Conclusions

Validity evidence supports use of the assessments by both educators measuring students’ CT practices and by researchers studying curriculum implementation and student learning in introductory high school computer science

Conclusions

Assessments are not “plug-n-play”

Test validation is not “one-and-done”

Each new use of the assessment requires ongoing investigation of the extent to which the available evidence supports the desired inferences one wants to make about test performance

More information

- More information about PACT?
  http://pact.sri.com/

- Review the ECS assessments and rubrics?*
  http://pact.sri.com/ecs-assessments.html

* Terms of Use & Licensing Information: https://pact.sri.com/assessment/termslicense.html

THE CS10K EVALUATOR WORKING GROUP

Rebecca Zarch

SAGEFOX CONSULTING GROUP

Evaluator Working Group (EWG)

The NSF CS10K program "aims to have rigorous, academic computing courses taught in 10,000 high schools by 10,000 well-prepared teachers."

How many teachers are being reached through the NSF- Funded CS10K program?

EWG Members

Rebecca Zarch	SageFox Consulting Group
Kathy Haynie	Haynie Research and Evaluation
Tom McKlin	The Findings Group
Christine Ong	UCLA, CRESST
Alan Peterfreund	SageFox Consulting Group
Gary Silverstein	Westat
Jeffrey Xavier	SageFox Consulting Group
Sarah Dunton	Expanding Computing Education Partnerships (ECEP)
Sarah Wille*	Outlier
Jenn Duck*	The Learning Partnership

*Prior Members

The challenge and approach

Context

- Multiple projects with unique models
  - *Including start and end dates*
- Each project with independent evaluation
- No mandate for participation

EWG approach

- Peer-driven approach
- Annual data spreadsheet
- Survey support document
- Community
  - *Validation*
  - *Feedback*
- Broader CS community

Guiding Questions

1. How many new teachers have participated in professional development (PD) through CS10K-funded projects?
   a. *What are the demographic characteristics of these teachers?*
   b. *What is their teaching experience?*
2. How many students have CS10K projects reached?
   a. *What are the characteristics of students that were reached through CS10K?*
   b. *What are the characteristics of the student subset who took the AP CSP exam?*
3. How many schools have a trained CS teacher?
   a. *What are the characteristics of the student body that has access to a course taught by a CS10K-trained teacher?*

Value of this approach:
Ability to say something about the capacity built nationally

- Teachers:
  - ***2,580 teachers*** *- **36** CS10K projects 2012-2016.*
  - ***Male (50%) White (79%)*** *and* ***non-Hispanic (90%).***
  - ***71% of teachers*** *with at least six years of K-12 teaching experience in any subject;* ***82% of teachers*** *were new to teaching computer science.*
- Students:
  - ***27,037 students*** *(At least)reached during the 2016-17 academic year; compared to 13,410 during the 2015-16 academic year..*
  - *More than* ***860,000 students*** *potentially have access to a CS10K teacher in 2016-17.*
    - 6% of the high school student population in the United States.
- Schools:
  - ***1,500 schools*** *across* ***45 states****, the District of Columbia, and Puerto Rico*
  - *In 2016-17,* ***778 schools*** *added newly trained CS10K CS teacher.*

What this approach misses?

Numbers don’t tell us …

- Quality of the training and/or curricular materials
- What happens in the classroom
- Teacher impact
- Student impact

Evaluation Wrecking Crew, CSONIC

Changing CS Ed landscape

- State policies (standards, credentialing, etc.)
- Multiple PD providers
- Multiple NSF funding mechanisms
  - *(e.g. CS10K, MSP, STEM+C, CSForALL RPP)*
- Multiple funding streams per project
  - *Public and private*
  - *Blended sources*

Next steps

- EWG
  - *Shifting focus to state/district data*
- Expanding Computing Education Pathways (ECEP)
  - *Meeting Jan 2018 -17 states considered feasibility of using state data*
  - *Includes state Department of Education representatives as partners*
- RPPforCS: Teacher PD
  - *Opportunity for systematic data from the start*
  - *Co-develop with the community*
  - *Researcher-Evaluator Working Group (R-EWG)*

Q & A

Acknowledgments

- The authors would like to acknowledge that this material is based upon work supported by the US National Science Foundation under Award Nos. DUE-1625335, DUE-1625005, CNS-1132232, CNS-1240625, DRL-1418149, CNS-1433065, CNS-1640237, and CNS-1228355

- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

CONTACT INFO

Adrienne Decker (adrienne.decker@rit.edu)

Monica M. McGill (mmmcgill@knox.edu)

Jason Ravitz (ravitz@google.com)

Eric Snow (eric.snow@sri.com)

Rebecca Zarch (rzarch@sagefoxgroup.com)