Universidad del Desarrollo

# The role of assessment in promoting social justice

3rd International Virtual Meeting:
Teaching, Learning & Assessment In Higher Education

Dylan Wiliam (@dylanwiliam)

# Outline

- What is assessment?
- What is assessment for?

# A framework for thinking about change

# There are no perfect solutions; only trade-offs

- Proposals for change are usually
  - clear about what the old approach did badly, and how the proposal will improve those aspects
  - silent about the things that the old approach did well, and that the new proposal will do less well
- Proposals for change should answer two questions:
  - "What will be better if the changes are made?"
  - "What will be worse if the changes are made?"
- If the answer to the second question is "nothing" then the proposer needs to think again.
- There will always be trade-offs
  - The question is whether they are explicit or not

# Assessment

# Assessment

- Assessment is a procedure for drawing inferences
  - We give learners things to do
  - We identify the evidence
  - We draw conclusions

# Evolution of the idea of validity

- A property of
  - a test
  - students’ scores on a test
  - inferences drawn on the basis of test results
- “One validates not a test but an interpretation of data arising from a specified procedure”(Cronbach, 1971)

# Prepositional permutations

- Assessment of learning
  - Status
  - Competence
  - Ranking
  - Prediction
  - Accountability
- Assessment for learning
  - Improving learning
- Assessment as learning
  - The learner’s role

Black, Harrison, Lee, Marshall and Wiliam (2004); Earl (2003)

# Assessment for (and as) learning

# Assessment for learning (Mittler, 1973)

- Assessment for learning is a broader concept than formative assessment
  - Assessment for motivation
  - Assessment for retrieval practice
    - Effective even when tests are not marked
    - Hypercorrection effect

# How can assessment improve learning?

| Announced? | Given? | Scored? | Used? | |
|---|---|---|---|---|
| | | | | Assessment for motivation |
| | | | | Retrieval practice |
| | | | | Instructional correctives |
| | | | | Formative assessment |

# Unpacking Formative Assessment

| | Where the learner is going | Where the learner is now | How to get the learner there |
|---|---|---|---|
| Teacher | Clarifying, sharing, and understanding learning intentions | Eliciting evidence | Providing feedback that moves learners forward |
| Peer | | Activating students as learning resources for one another | |
| Student | | Activating students as owners of their own learning | |

# Unpacking Formative Assessment

| | Where the learner is going | Where the learner is now | How to get the learner there |
|---|---|---|---|
| Teacher | Clarifying, sharing, and understanding learning intentions | Eliciting evidence | Providing feedback that moves learners forward |
| Peer | | Activating students as resources for one another | |
| Student | | Activating students as owners of their own learning | |

# Learning intentions and success criteria

- Learning intentions
  - descriptions of intended *learning*
  - useful for *planning*
  - mainly useful for teachers
- Success criteria
  - descriptions of *performance* on learning tasks
  - primarily useful for *evaluating*
  - useful for both teachers and students

# Eliciting evidence

- Two good reasons to ask a question
  - Cause thinking
  - Provide data that informs instruction
- Better evidence
  - Deeper
  - Broader
- Creating, capitalizing on *moments of contingency*

# Feedback

- The purpose of feedback is to improve the student, not the work
- The only thing that matters with feedback is what students do with it
- If your feedback is getting you more of what you want, it’s good feedback
- Feedback should be more work for the recipient than the donor

# Cooperative and collaborative learning

- Key issue: who decides the goal?
  - Learners: “collaborative learning”
  - Teachers: “cooperative learning”
- The purpose of collaboration and cooperation
  - To produce the best solution
  - To improve individuals’ ability to collaborate
  - To maximize learning for the *group*
  - To maximize learning for *each member* of the group

# Students as resources for one another

- Group goals:
  - so students are working *as* a group, not just in a group
- Individual accountability:
  - the best learning efforts of every member of the group must be necessary for the group to succeed, and
  - the performance of each group member must be clearly visible and quantifiable to the other group members

Slavin, Hurley and Chamberlain (2003)

# Students as owners of their own learning

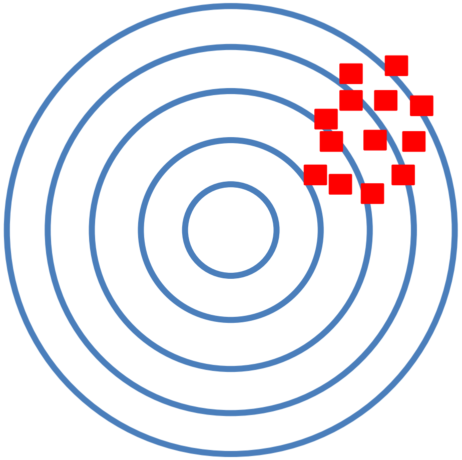
- Students assessing their own work:
  - With rubrics
  - With exemplars
- Self-assessment of understanding:
  - Learning portfolio
  - Plus/minus/interesting
  - Practice testing

# Assessment of learning

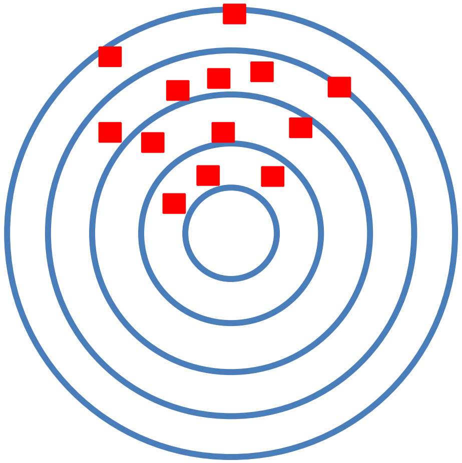
# Reliability

- Validity and reliability?

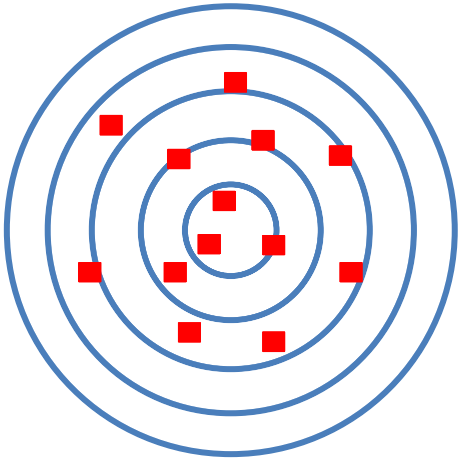
# A common—but limited—metaphor



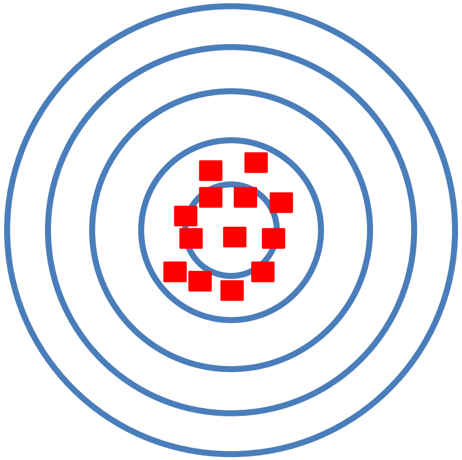
Reliable but not valid



Neither reliable nor valid



Valid but not reliable



Both valid and reliable

# Reliability

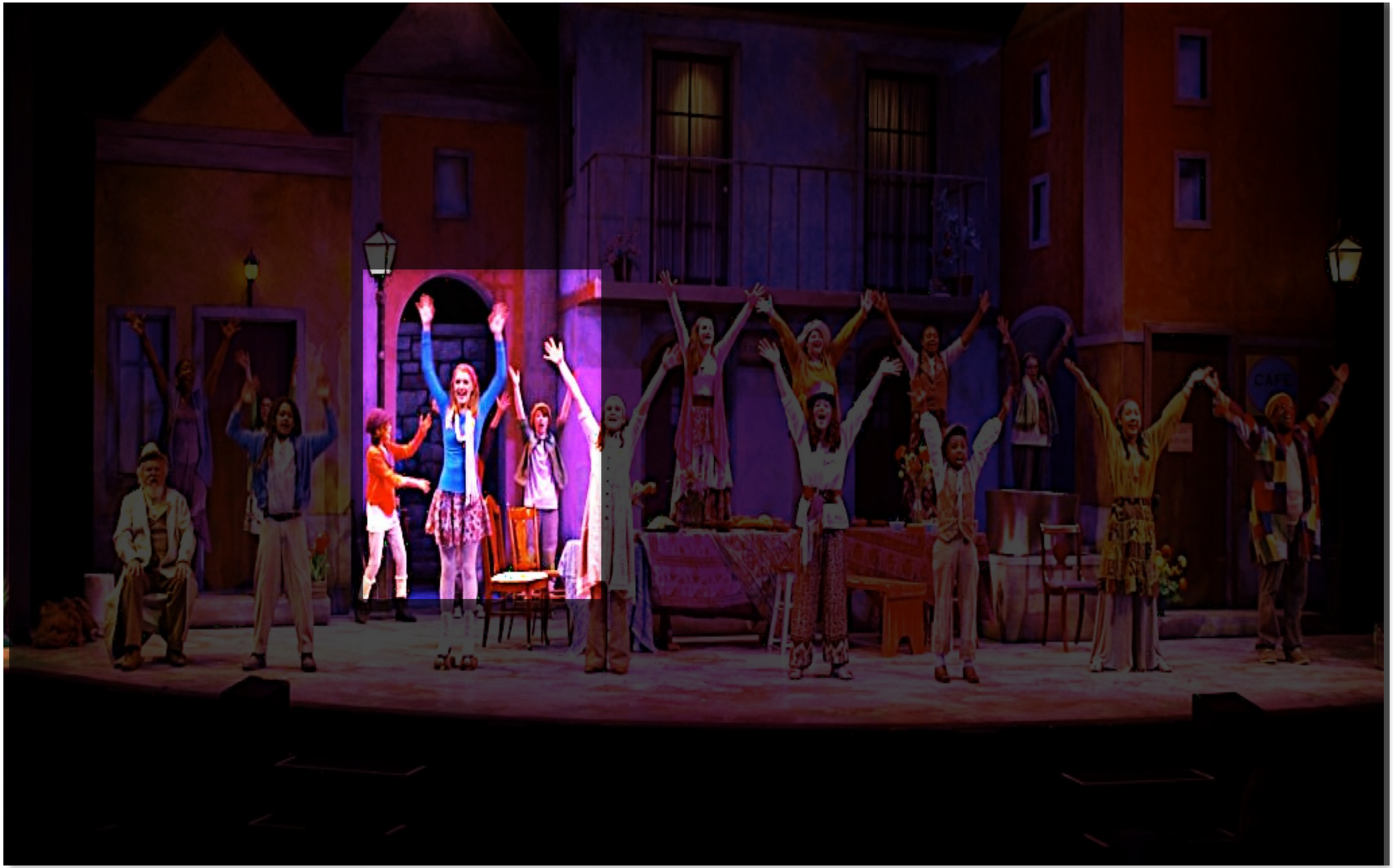
- Validity and reliability?
- Validity versus reliability?

# Resolving the validity-reliability paradox

- The paradox in brief
  - Reliability is a pre-requisite for validity
    - Unreliable scores cannot support valid inferences
  - Reliability is the enemy of validity
    - Attempts to increase reliability generally reduce validity
- The resolution
  - Increasing reliability
    - strengthens some inferences
    - weakens other inferences

# The “stage lighting” metaphor: Floodlight

# The “stage lighting” metaphor: Spotlight



# Reliability

- Validity and reliability?
- Validity versus reliability?
- Validity *including* reliability

# Two main threats to valid interpretations

## Assessment is ‘too big’

- Scores depend on things they shouldn’t
  - Irrelevant factors
    - Handwriting
  - Luck
    - Good/bad days
    - Who marks the work
    - The particular tasks set
- Construct-irrelevant variance

## Assessment is ‘too small’

- Scores don’t depend on things they should
  - Assessing only things that are easy to assess
- Construct under-representation

# Assessment

- Implications
  - No such thing as a valid assessment
  - Validity is a property of inferences
  - Reliability is part of validity
    - Specifically, the random component of construct-irrelevant variance
  - No such thing as a biased assessment
    - A test tests what a test tests
  - No such thing as a formative assessment
    - the terms f*ormative* and *summative* are properties of inferences, not assessments

# Constructs and equity: Examples

# Assessing history

- On multiple-choice tests of history, males outperform females
- On constructed-response tests of history, females outperform males
- Common interpretations
  a. Multiple-choice tests are biased against females
  b. Constructed-response tests are biased against males

# Construct definition

- “knowing facts and dates”
  - Multiple-choice tests are ideal because it is possible to assess many facts and dates
  - Constructed response tests assess things they shouldn’t, like language skills, handwriting
  - Constructed response tests ask fewer questions, this increasing the role of luck

- “describing and explaining historical events”
  - Multiple-choice tests are inadequate (scores don’t depend on things they should)
  - Constructed-response tests are essential (so that scores depend on things they should)

# Performance assessment

- Key requirements
  - assessment is direct rather than by correlates
  - demonstration of task performance
- Examples
  - Driving test
  - Objective Structured Clinical Examination
- Optional add-ons
  - authentic
  - meaningful to the student
  - engaging
  - applied

# Trade-offs in performance assessment

- Pros
  - Assess capabilities that cannot be assessed in other ways
  - Engaging and motivating for students
  - Face validity

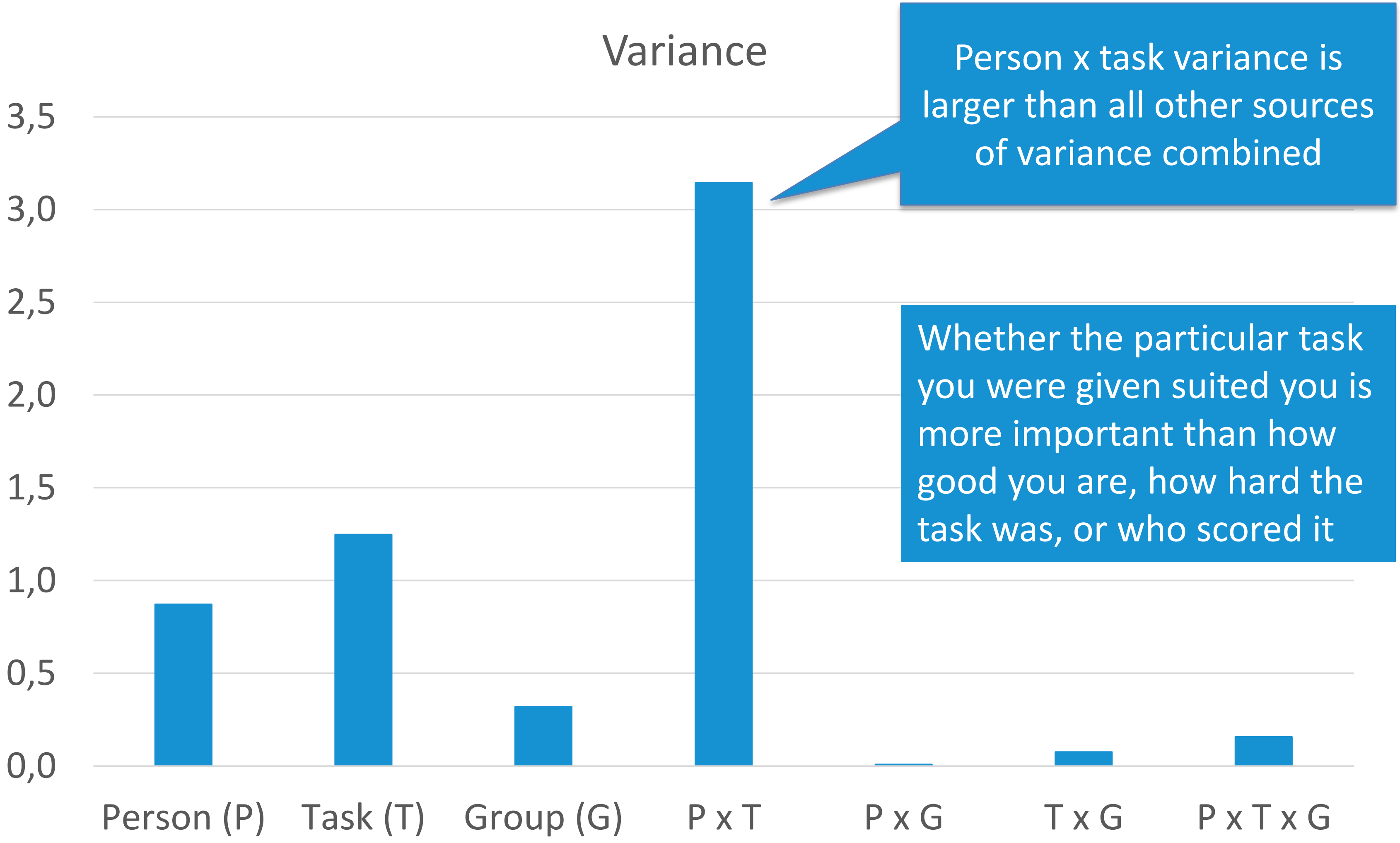
- Cons
  - Construct definition is challenging
  - Comparability
  - Disclosure
  - Generalizability
    - Scoring
    - Student x task interaction

# Case study: Assessing patient management skills

- 200 medical students randomly selected from a larger group that had completed a performance assessment of patient management skills online
- Three groups of four university faculty primary care physicians assessed 100 of the students’ responses
- Results used to create three scoring algorithms for each task
- Algorithms used to score the responses of the other 100 students

Clauser, Harik, Clyman (2000)

# Results



Person x task variance is larger than all other sources of variance combined

Whether the particular task you were given suited you is more important than how good you are, how hard the task was, or who scored it

Clauser, Harik, Clyman (2000)

# Trade-offs in assessment design

- Distributed
  - So that evidence collection is not undertaken entirely at the end
- Synoptic
  - So that learning has to accumulate
- Extensive
  - So that all important aspects are covered (breadth and depth)
- Manageable
  - So that costs are proportionate to benefits
- Trusted
  - So that stakeholders have faith in the outcomes

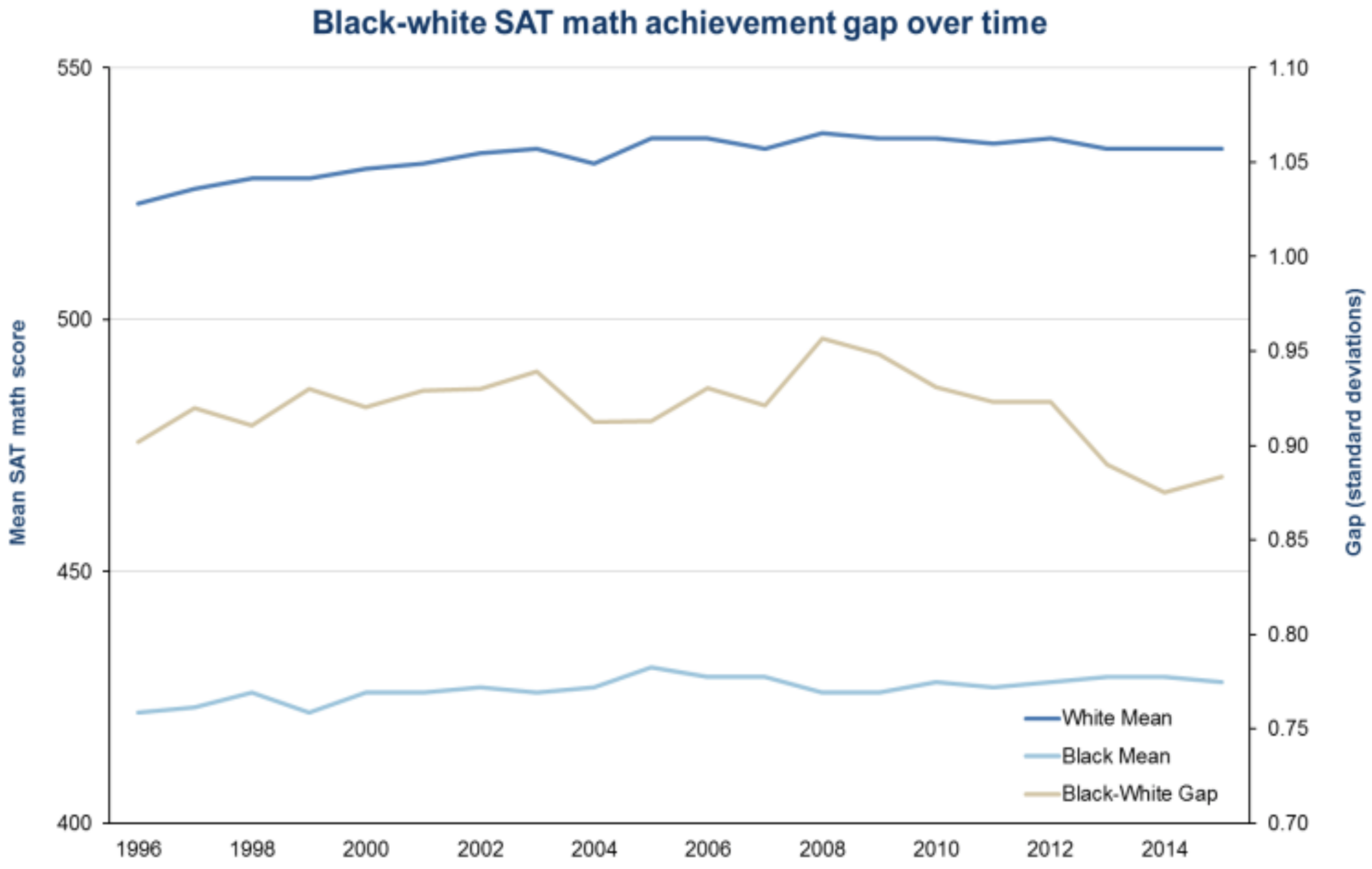
# Selection for higher education

# Selection for higher education: The SAT

- “The most researched test in the world” (College Board, 2009)
- Correlation with
  - First year GPA: 0.44 to 0.62 (Hezlett et al., 2001)
  - Cumulative GPA: 0.35 to 0.45 (Hezlett et al., 2001)
  - Comparable to high-school GPA (Kobrin, Camara, & Milewski, 2002)
- Even better when combined with HSGPA

| | |
|---|---|
| – Hispanic students | 0.42 |
| – White students | 0.48 |
| – Asian American students | 0.55 |
| – African American students: | 0.55 |

# Is the SAT fair?



# Potential causes

- Sample differences
- Construct-irrelevant variance
  - “the SAT has been shown to be both culturally and statistically biased against African Americans, Hispanic Americans, and Asian Americans” (Freedle, 2003 p.1)
- Construct-*relevant* variance

# Prediction of 1st year GPA in 23 colleges

| | HSGPA | | SAT | | HSGPA+SAT | |
|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male |
| African American | 0.09 | **0.20** | 0.01 | **0.22** | -0.01 | **0.16** |
| Native American | **0.13** | **0.24** | 0.06 | **0.32** | 0.07 | **0.28** |
| Asian American | 0.03 | 0.08 | -0.01 | **0.15** | -0.03 | 0.07 |
| Hispanic | **0.23** | **0.31** | 0.03 | **0.20** | 0.04 | **0.20** |
| White | **-0.11** | -0.03 | -0.11 | 0.06 | -0.09 | 0.05 |
| Other | -0.09 | 0.04 | -0.13 | 0.03 | **-0.12** | 0.04 |

Kobrin, Camara, and Wilewski (2002)

# Construct-relevance depends on constructs!

- What is the purpose of the SAT?
  - To predict student performance on existing college programs?
  - To predict student ability to thrive in higher education?
  - To indicate which students may need additional support to thrive?

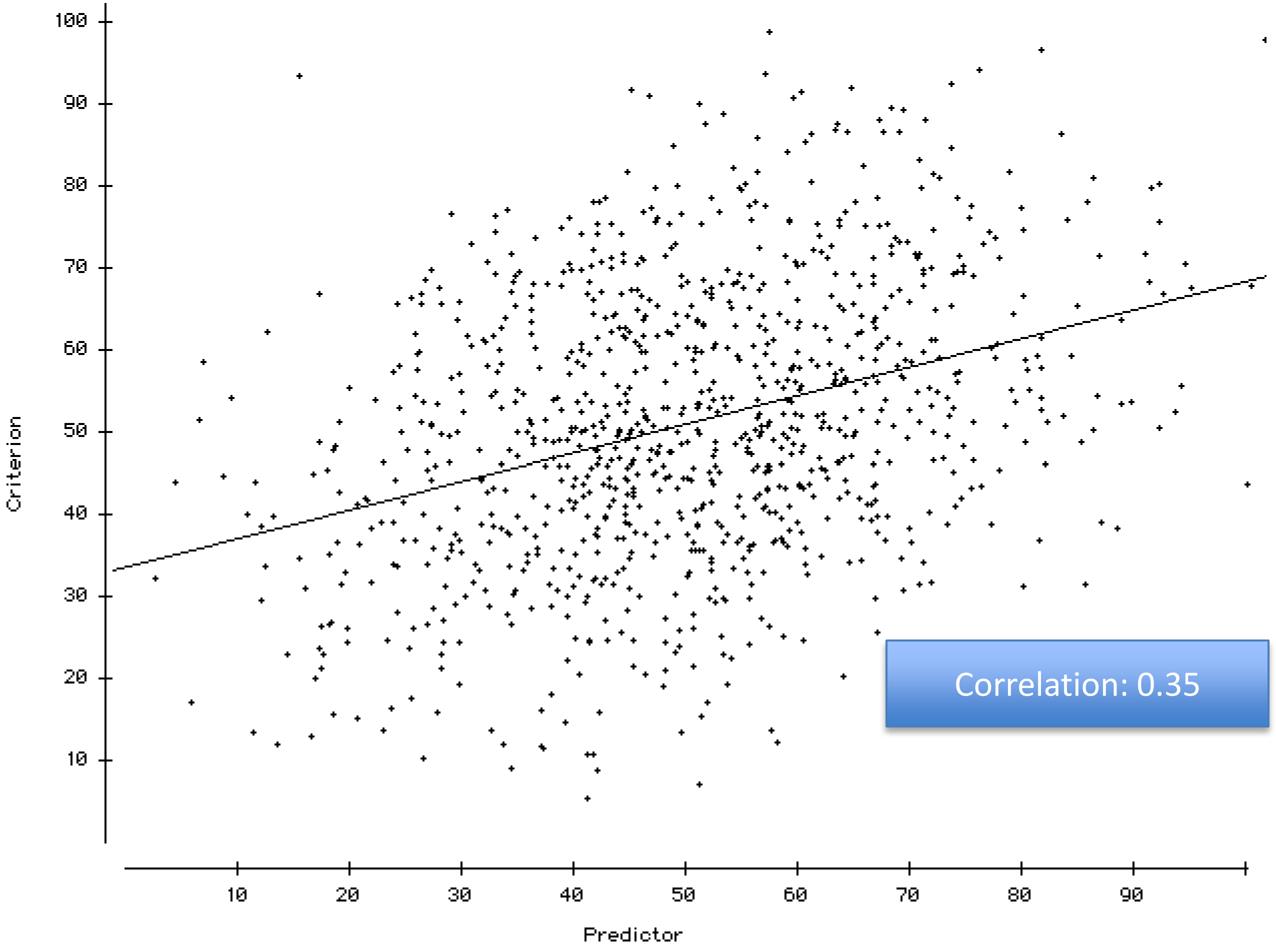
# An illustrative case study: Access to Medicine

# Access to Medicine: the context

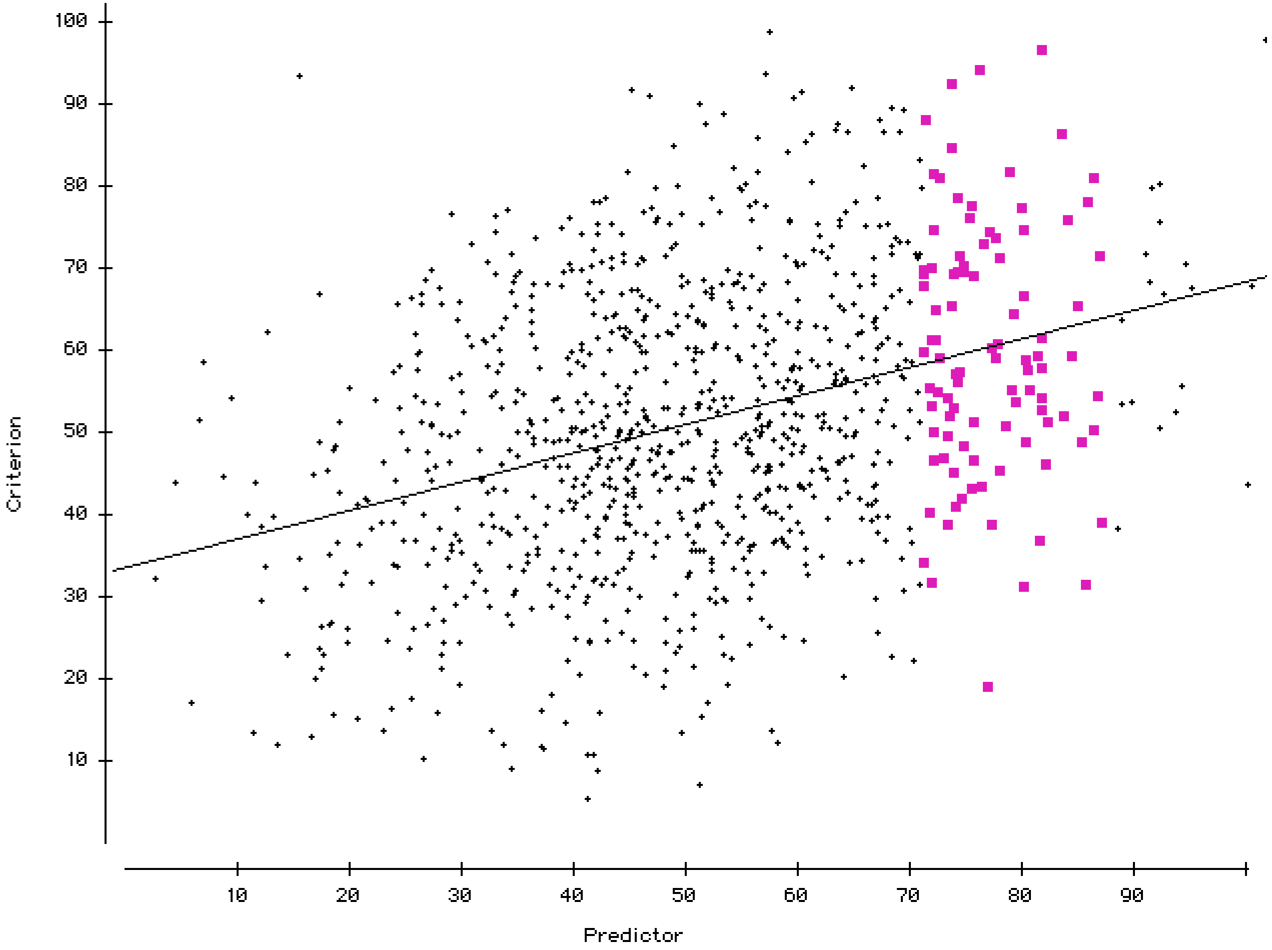
- King’s College London
  - 5th largest university in the United Kingdom
  - Largest medical school in Europe
  - Located in SE London: ethnically diverse population
  - Highly selective admissions, based on achievement on high-stakes examinations
  - Result:
    - Private school students over-represented
    - White and Asian students over-represented
    - Students of African heritage under-represented
    - Socio-economically disadvantaged students under-represented

# Access to Medicine

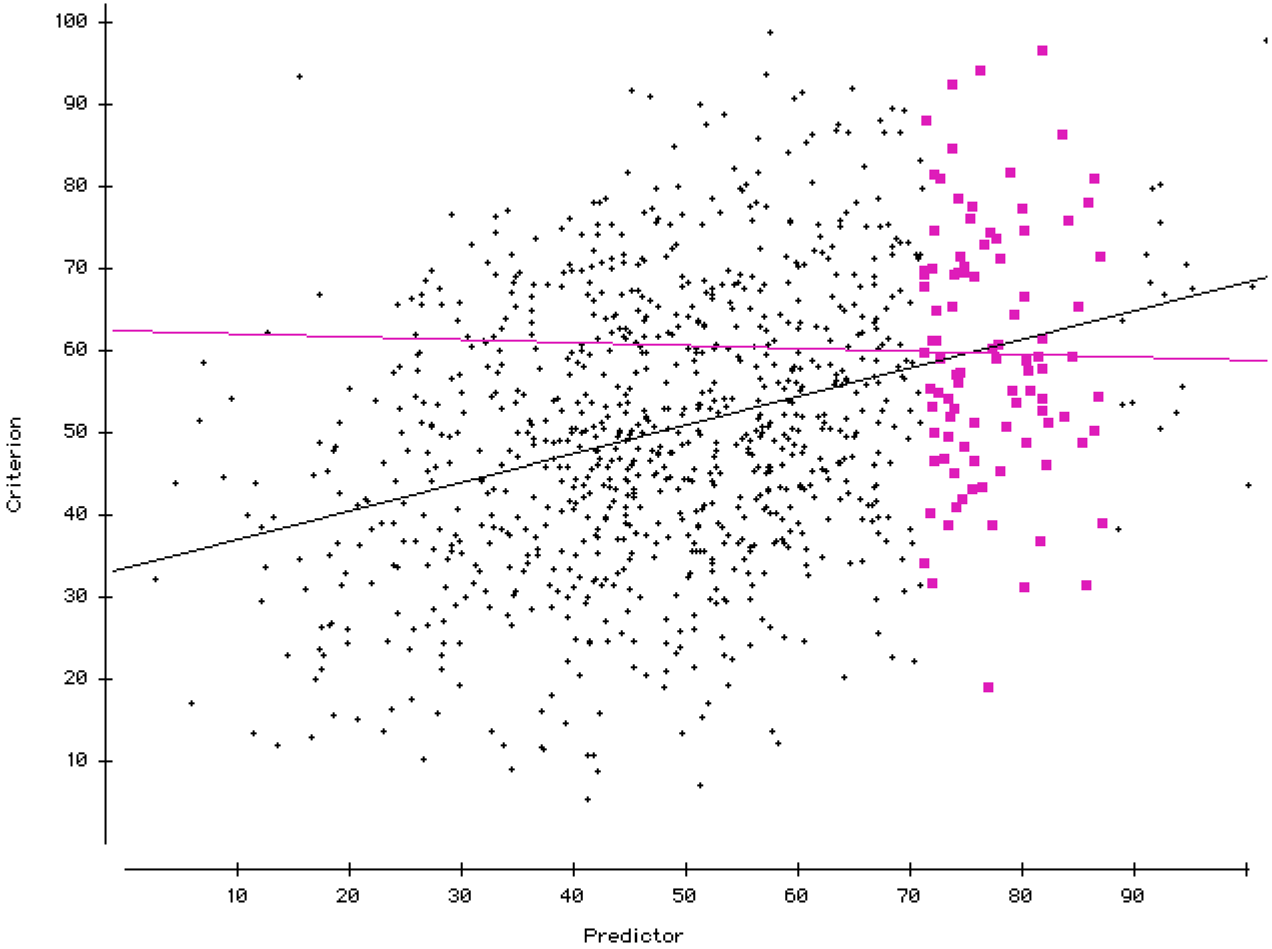
- The goal:
  - Widen access to, and success in, medical education
  - Produce more “culturally competent” medical services
- The challenge
  - Do so without “lowering the bar”
  - Do not provide a “second chance” for already advantaged students



Correlation: 0.35



100
90
80
70
60
50
40
30
20
10
Criterion
10
20
30
40
50
60
70
80
90
Predictor


100
90
80
70
60
50
40
30
20
10
Criterion
10
20
30
40
50
60
70
80
90
Predictor

# A solution in three parts…

- Recruitment: Program to raise aspirations in middle school
- Selection: Only public-school students in local area eligible
  - Assessment via school achievement, personal characteristics and science reasoning tasks
- Retention
  - Additional foundation year (two years to cover traditional first year curriculum)
  - Dedicated support
    - living expenses and tuition provided for first year only
    - dedicated tutorial support

# Science Reasoning Tasks (SRTs)

- Based on work of Shayer & Adey (1981)
- Suite of group-administered tasks
- Assess not science knowledge but ability to incorporate new facts into existing schema
- Benchmarked on *existing medical students*
- So provide *alternative ways* of showing talent

# The story so far

- Program in its 16th year (now “Extended Medical Degree Programme”
- Steady state: 300 students enrolled (50 in each year)
- Still substantial challenges
  - High maintenance cost
  - Selection methods modified over time
- But
  - Non-traditional students indistinguishable from traditional route students
  - Not seen as a ‘soft option’

# Validity revisited

“Validity is an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.” (Messick, 1989 p. 13)

- Social consequences:
  - “Right concern, wrong concept” (Popham, 1997)
  - No such thing as “consequential validity”

- “As has been stressed several times already, it is not that adverse social consequences of test use render the use invalid, but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct-irrelevant variance.” (Messick, 1989 p. 88)

- “If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced—or if the validation process can discount sources of test invalidity as the likely determinants, or at least render [page break] them less plausible—then the validity of the test use is not overturned. (p. 88-89)

- "Adverse social consequences associated with valid test interpretation and use may implicate the attributes validly assessed, to be sure, as they function under the existing social conditions of the applied setting, but they are not in themselves indicative of invalidity." (p. 89)

# Summary (1)

- Validity is a property of inferences, not of assessments
- Two threats to valid inferences
  - Scores depend on things they shouldn't
    - construct-irrelevant variance
  - Scores don't depend on things they should
    - construct underrepresentation
- Construct definition is essential because
  - With good construct definition
    - assessment design is a technical process, so
    - construct definers are in charge
  - With poor construct definition
    - assessment design is a values-laden process, so
    - assessment developers are in charge

# Summary (2)

- The main issue: What is the source of any differential impact?
  - “Unpacking” the construct-irrelevant variance
    - Random: Does the result depend on chance factors?
    - Systematic: Does the assessment support the same inferences for all groups?
  - Construct relevant variance
    - Is it the right construct?
    - Would a different construct produce a more equitable outcome?

# Thank You

www.dylanwiliam.net