Socioculturally Responsive Assessment: What is it and What Does it Look Like?

Randy Bennett

Educational Testing Service
Princeton, NJ 08541
rbennett@ets.org

Virtual presentation at the Learning Sciences Research Institute and Department of Educational Psychology, University of Illinois Chicago, April 2022

Three-Part Series

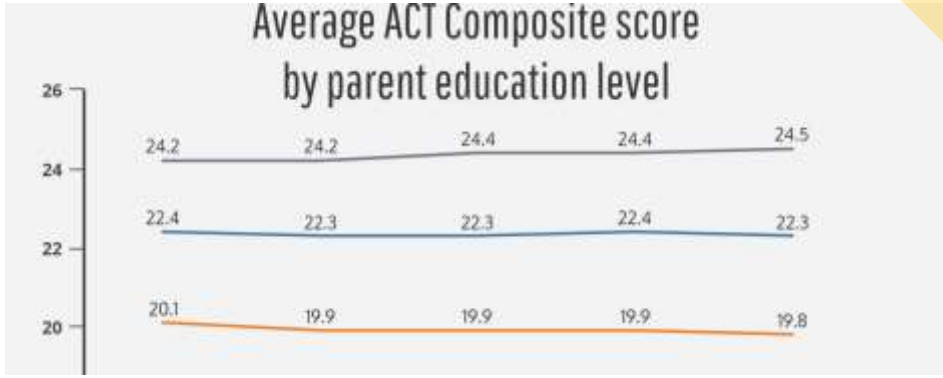
I: Equity and Assessment in the Post-COVID-19 Era

II: The Good Side of COVID-19

III: Socioculturally Responsive Assessment: What is it and What Does it Look Like?

I: Equity and Assessment in the Post-COVID-19 Era

- Structural inequity is longstanding and pervasive
- Limits OTL
  - School, home, community
  - Over generations
- Structural inequity affects accomplishment
  - Limits prior knowledge facilitative for school
  - Contributes to large group differences in test scores, grades, and life chances
  - Influences what's valued by the school



II: The Good Side of COVID-19

- US rooted in a dominant common culture
  - All groups have not had equal opportunity to participate
- Testing born of, and helps to effect membership in, that culture
- US rapidly becoming multicultural
- Population dynamics make unsustainable:
  - Continuation of a common culture as it now operates
  - Testing's traditional role
- COVID-19 helped fuse testing to social injustice

Educational Measurement
ISSUES AND PRACTICE
NCME

Educational Measurement: Issues and Practice
January 2022, Vol. 0, No. 0, pp. 1–3

The Good Side of COVID-19

Randy E. Bennett, *Educational Testing Service*

Abstract: *This commentary focuses on one of the positive impacts of COVID-19, which was to tie societal inequity to testing in a manner that could motivate the reimagining of our field. That reimagining needs to account for our nation's dramatically changing demographics so that assessment generally, and standardized testing specifically, better fit the needs of a multicultural society.*

Keywords: assessment, COVID-19, equity, fairness, testing

COVID-19 did many things to our world, most of which ranged from bad to terrible. It did do some good things, however, one of which was to make more immediately visible the fundamental structural inequities that characterize virtually every aspect of life in our country (Aspen Institute Roundtable on Community Change, 2004). A second positive impact was to accelerate debate over the appropriate role of traditional single-event, standardized tests for school accountability and for university admissions. A third salutary impact was to fuse inequity with testing in a way that might help us begin to refashion our field. In this commentary, I will explicate that last claim and some of its implications.

It takes little insight to observe that our country is experiencing a political divide of greater depth and magnitude than seen in decades. One way of framing that divide is in terms of differing visions for the future, each of which has significant implications for testing (and assessment more broadly). On one side of that divide is a historical conception of America as the metaphoric "Melting Pot" (Hanson, 2016). In that vision, those who were culturally different from the country's Western European founders progressively took on the dominant group's values, beliefs, knowledge, and practices. Through assimilation, and with ability and effort, these former outsiders meritocratically moved toward the "American Dream" (Adams, 1931)—a good job, a middle-class lifestyle, a home, and a nest egg to be used for their children's college education, for their own retirement, or as a legacy. Whereas assimilation was often voluntary, it was in other cases coerced (e.g., conscription of native American children to boarding schools; Pember, 2019) and, in still others, actively impeded via structural barriers (e.g., slavery, Federal Housing Administration redlining of African American neighborhoods).

One of the primary mechanisms of such assimilation is education. Public schools are designed to inculcate children into a common culture—or generally similar versions of it—as represented in state content standards and in community values. Traditionally, this inculcation has meant teaching the Pledge of Allegiance and Star-Spangled Banner, standard American English, a subset of world literatures representing English and (Caucasian) American authors, and history emanating from a Western European perspective aligned, for example, with traditions like Columbus Day (a federal holiday).

Standardized testing is an important component of education's assimilative function. State accountability tests help policymakers, educators, parents, and the public understand how effectively schools are imparting certain aspects of the common culture. For admission to selective postsecondary programs, standardized tests have been the meritocratic machinery that traditionally helped allocate the opportunity to immerse oneself further in that common knowledge-and-value set, climbing higher on the ladder to the American Dream.

The appropriateness of using tests for evaluating schools and for awarding admission to selective higher education institutions has been a topic of debate for a considerable time. COVID-19, however, gave policymakers license to act. School accountability testing was cancelled in spring 2020 and returned in considerably amended form in 2021 (e.g., fewer questions, remote delivery, relaxed participation requirements). More profoundly, earlier debate over whether the common culture represented in school curricula was itself inclusive enough (e.g., Paris & Alim, 2014), began to find voice in the measurement community (Randall, 2021). That inclusivity is conceptualized in several ways—e.g., better representing the contributions and cultures of diverse groups, connecting to the unique "funds of knowledge" common to different cultural communities, and more honestly depicting the history and continuing presence of structural inequity and racism in our society.

One of the reasons that COVID-19 was successful in advancing such equity issues is a matter of demographics. Changes in the U.S. population make the trend clear: In 1980, the population was 80% non-Hispanic White (Hobbs & Stoops, 2002, p. A35, table 10); by 2020, that figure had declined to 58% (Schneider, 2021). Public school demographics are even more telling: As of 2018, our school population was 47% non-Hispanic White (NCES, 2020, table 203.70) and, in our most populous state (California), the comparable figure in 2019–2020 was 22% (CDE, 2021). These demographics make debatable at best the continued presumption of a common culture based on a (Caucasian) Western European ideal.

One might argue, then, that we are witnessing a contention between two visions for our country's future—that of the Melting Pot versus a more multicultural one. In

Bottom Line

- Change our:
  - Mindset
  - Theories
  - Practices
  - Tools
  - Interpretations and uses
- *Thoughtfully* adapt to the more multicultural, pluralistic society the US is becoming

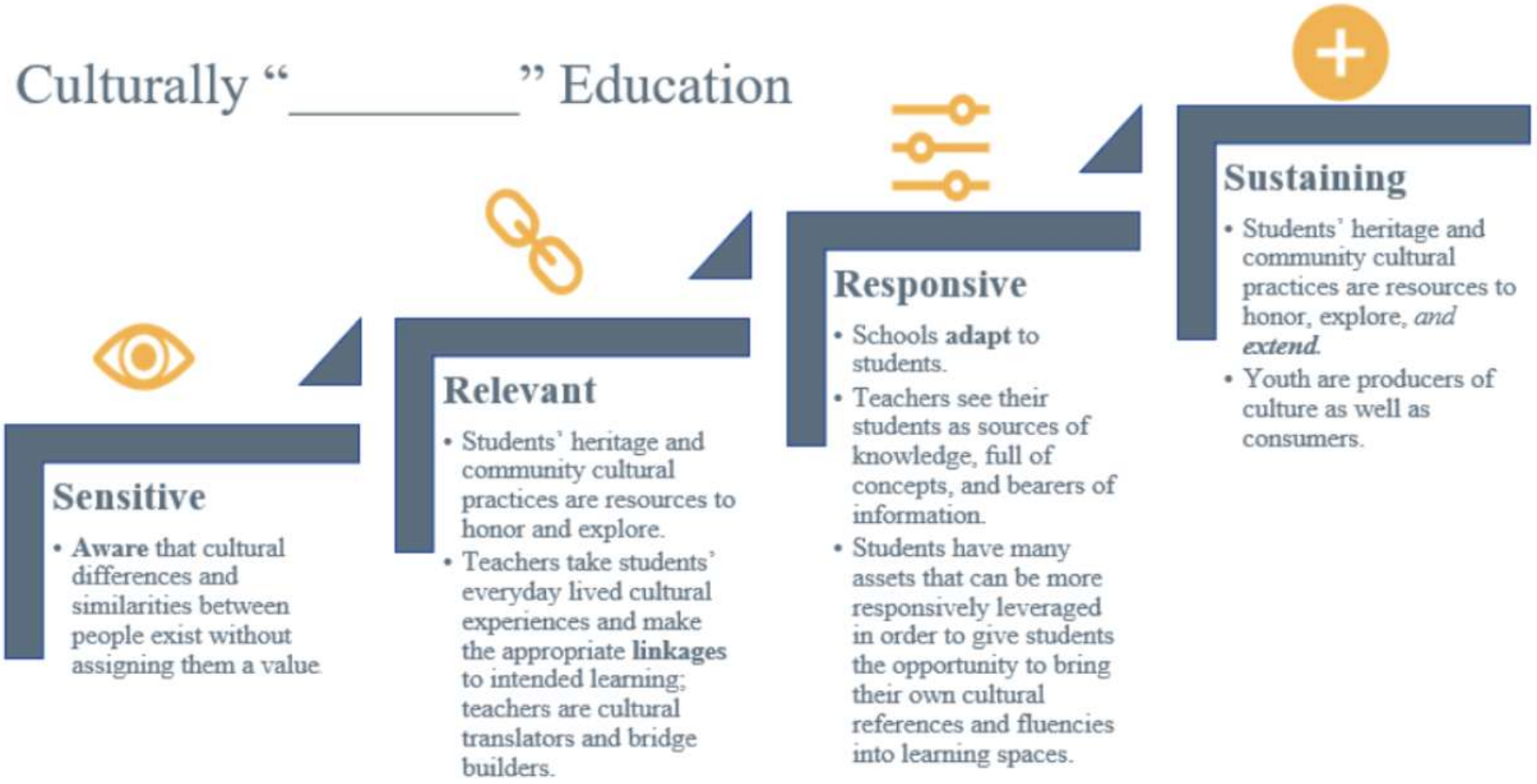
Overview

- Socioculturally responsive assessment
  - The end goal
  - Provisional principles for assessment design
  - A working definition
  - An initial theory
  - A suggested path to implementation
  - Summary

The End Goal

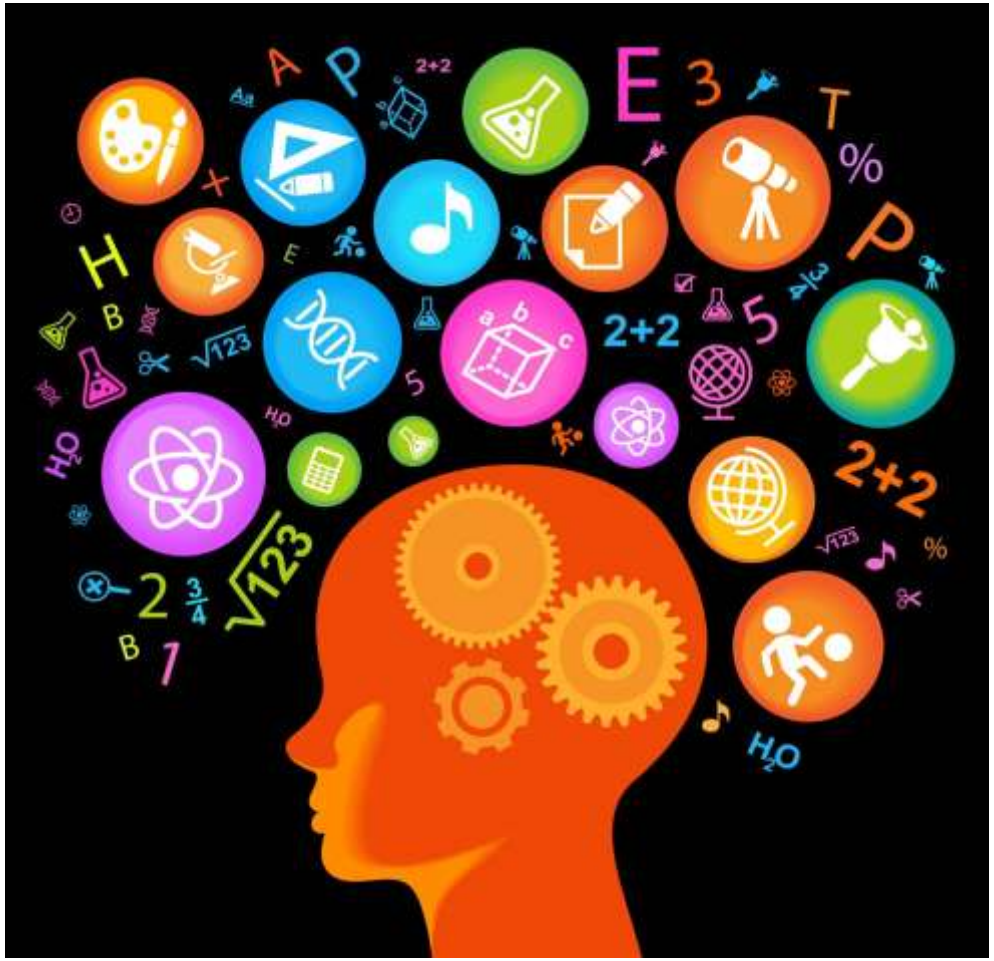
- Assessments "*born* socioculturally responsive"
  - Derivations
    - "Born accessible" (UDL)
      - Content structured to allow easy interaction for individuals with disabilities
    - Culturally relevant pedagogy
    - Socioculturally responsive classroom assessment
  - Key idea
    - Design for the social and cultural characteristics of students and the contexts from which they come

Hierarchical Model (Evans, 2021)



Principle 1: *Present problem situations that connect to, and value, examinee experience, culture, and identity*

- Rationale
  - Students more likely to show what they know in familiar vs. foreign contexts
    - Theoretical basis: Culturally relevant pedagogy, “funds of knowledge,” learning sciences (Ladson-Billings, 1995; Moll, 2019)
      - Connect academic content and processes valued by school to prior knowledge diverse learners bring from home and community
      - Recognizes importance of academic competency and of underutilized assets students possess



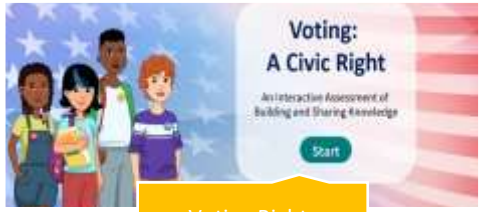
Principle 1: *Present problem situations that connect to, and value, examinee experience, culture, and identity* (con't)

- Rationale
  - Communicate respect for different cultural identities to help sustain them and help others learn about them (Paris, 2012)
  - Broaden the content and processes valued by the school
    - Consistent with state frameworks and curriculum requirements for:
      - Culturally responsive teaching
      - Native American history and culture
      - Asian American and Pacific Islander history
      - Ethnic studies

Example: ETS Testlets

Historical Voices

United Farm Workers



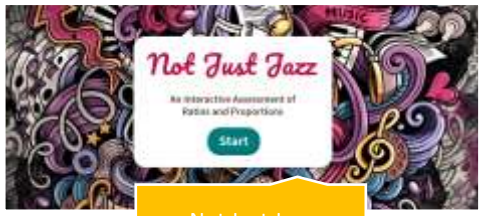
Voting Rights

Urban Gardens

Day of Service

Community Murals

Poetry on the Bus



Not Just Jazz

Print It

Hillside Farming

Wish a Wig

Greatest of All Time

Manga

Culture Fair

Breaking Barriers

Building an Amphitheater

Example: Present problem situations that connect to, and value, examinee experience, culture and identity

Immigrant groups in the US bring significant assets and encounter significant challenges that may be associated with sociocultural characteristics (e.g., language, beliefs and practices, parents' education level, economic status, race/ethnicity). In a written response:

- Describe a sociocultural characteristic of one or more immigrant groups
- Identify a significant asset associated with that characteristic, as well as a way in which societal structures might pose challenges for individuals with that characteristic
- Discuss how that asset and challenge might affect the school experience of learners from that group
- Suggest how teachers and student peers might use that asset to help those students progress

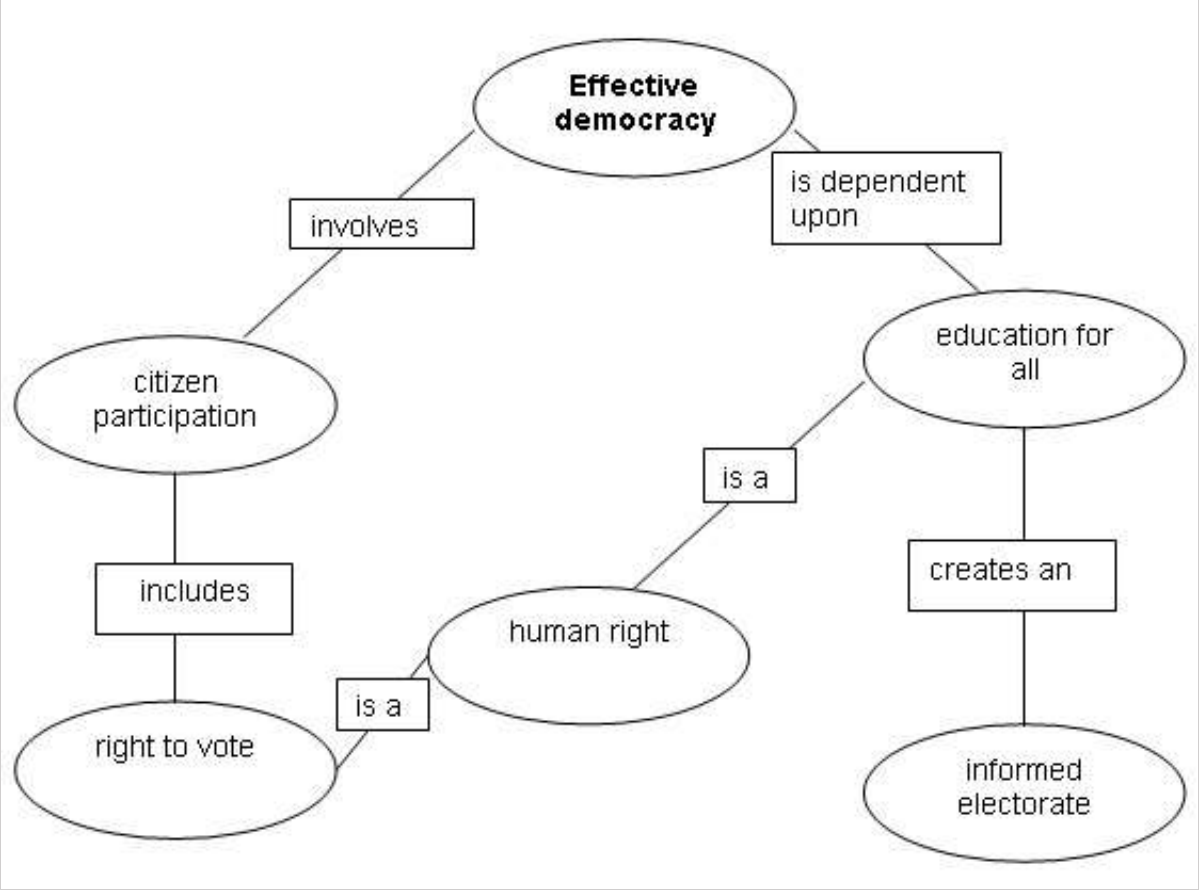
Note. Adapted from California Commission on Teacher Credentialing, *California Teacher of English Learners Study Guide* (Section 5), 2021, Retrieved from https://www.ctcexams.nesinc.com/content/docs/CX_SGsection5.pdf

Principle 2: *Allow for multiple forms of representation and expression in problem stimuli and in responses*

- Rationale
  - Some forms may be more common to the ways of knowing and community practices of particular groups
    - Braille/large type/audio for visually impaired examinees
  - Argues for greater allowance of:
    - Verbal, graphical, and symbolic representations
      - See Mayer (2009) for guidance on simultaneous use
    - Oral and written expression

Example: *Allow for multiple forms of representation and expression in problem stimuli and in responses*

What does it take to make democracy work? Write your response in a few sentences, a hierarchical list, or a concept map.

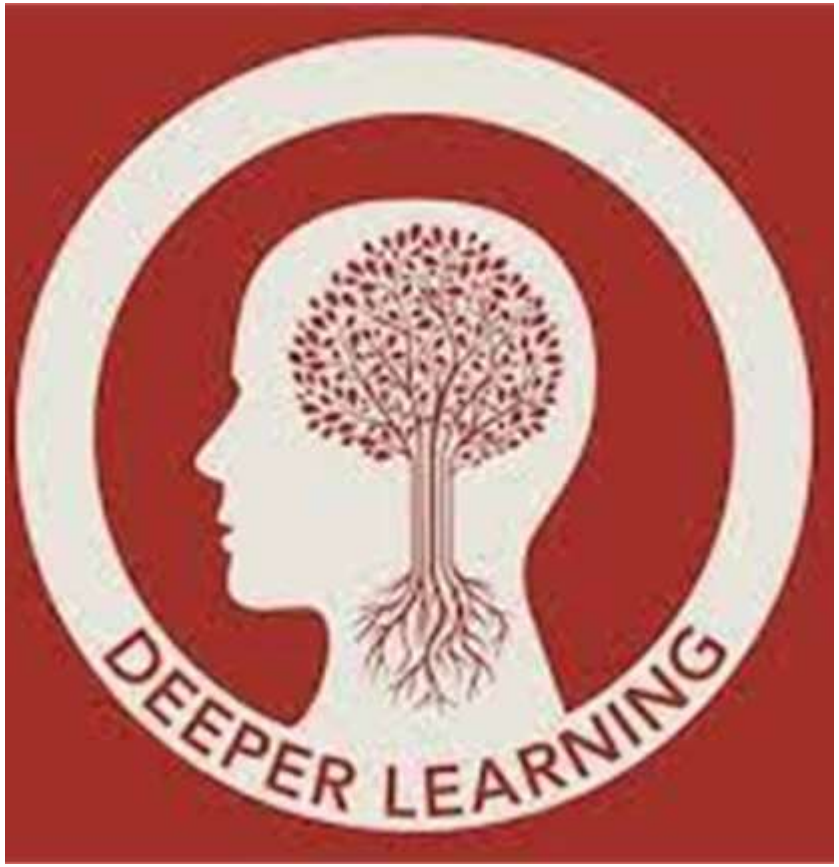


Principle 3: *Promote instruction for deeper learning through assessment design*

- Three Premises
  - Conceptual understanding critical and greater attention needed (NRC, 2000)
  - High-performing schools and middle-class homes provide deeper learning routinely
    - Students from diverse groups taught by less experienced/ qualified teachers with less access to curricular resources
  - Assessment design influences teaching and learning behavior
    - MCQ => instruction oriented to disconnected facts and memorized procedures
    - Performance tasks encourage teaching conceptual understanding, knowledge application, and skill integration (Stecher, 2010)

Assessment Design Requirements for Deeper Learning

- Include performance tasks posing reasonably realistic problems
- Provide consultative resources
  - K-12 ELA: write an essay based on given sources
  - Occupational/professional:
    - ARE: make evaluative judgments based on multiple resources
- Model strategies proficient performers use



Example: Promote instruction for deeper learning through assessment design

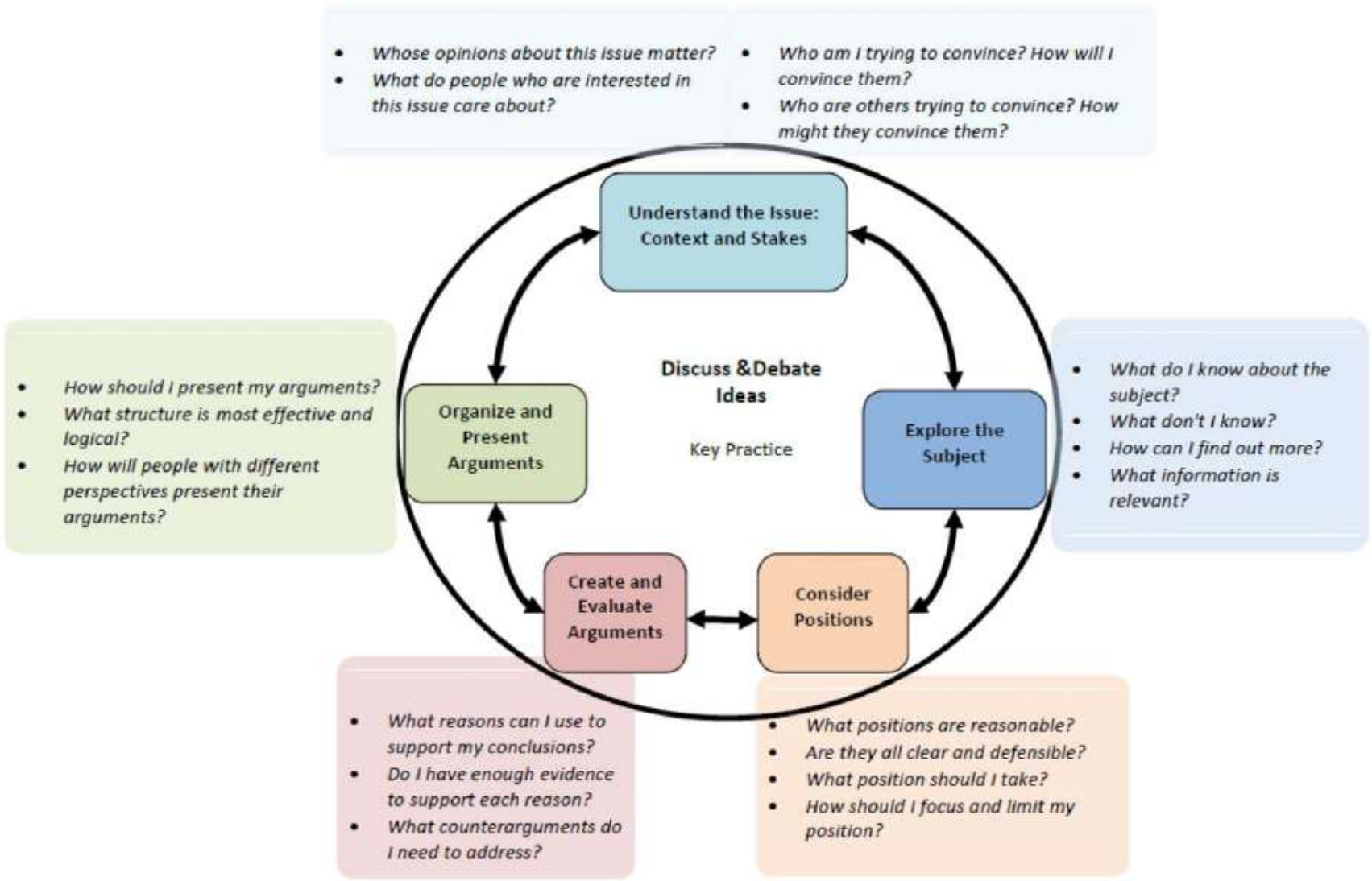


FIGURE 2 Major activity phases and associated goals for the key practice, discuss and debate ideas. From Dean and Song (2014). © 2014 Colegio Oficial de Psicólogos de Madrid. Reproduced by permission of Elsevier España.

Assessment Design Requirements for Deeper Learning

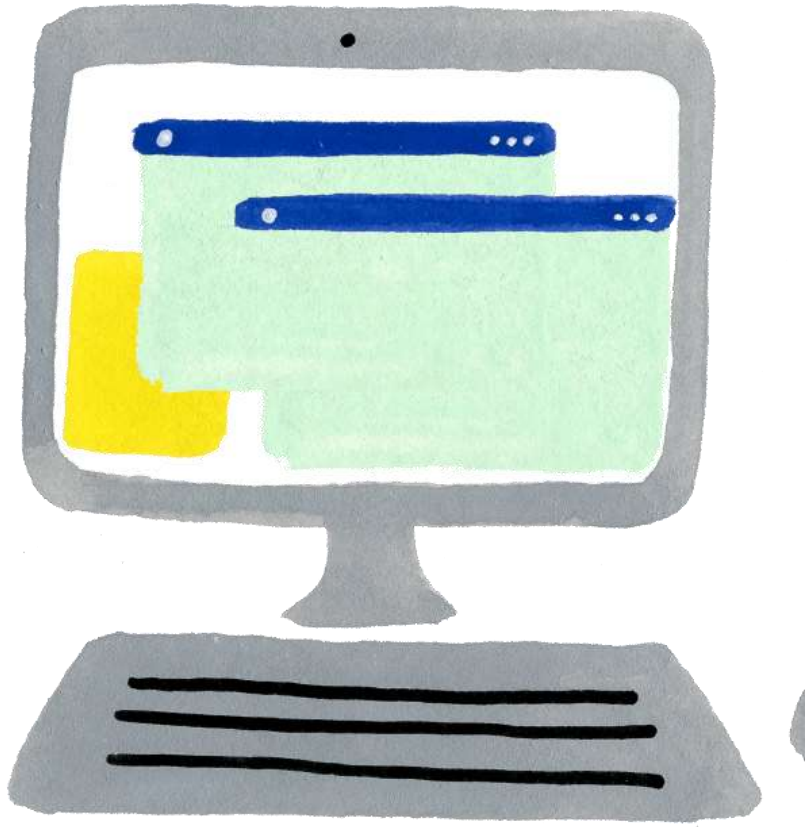
- Student agency
  - Increases motivation to learn (NASEM 2018; Shepard, 2021)
  - May help students learn to choose wisely
  - Goes beyond selecting response mode (Principle 2)
  - Encompasses problem choice

Principle 4: *Adapt the assessment to student characteristics*

- Testing premised on treating everyone the same
  - Content
  - Task format
  - Conditions
- Foundation for comparable scores
- Departures reduce score comparability and fairness

Adaptation is Not New

- Principled departures
  - Personalization on competency level
    - Stanford-Binet Intelligence Scales
    - Adaptive testing (1980s)
    - Common in state accountability and graduate/professional admissions
    - Examinees get different items, different difficulty
  - Personalization to special needs
    - 1937 creation of the Braille SAT
    - Accommodations now offered and treated as comparable
    - Some available to all examinees (English glossary, highlighter, strikethrough, zoom)

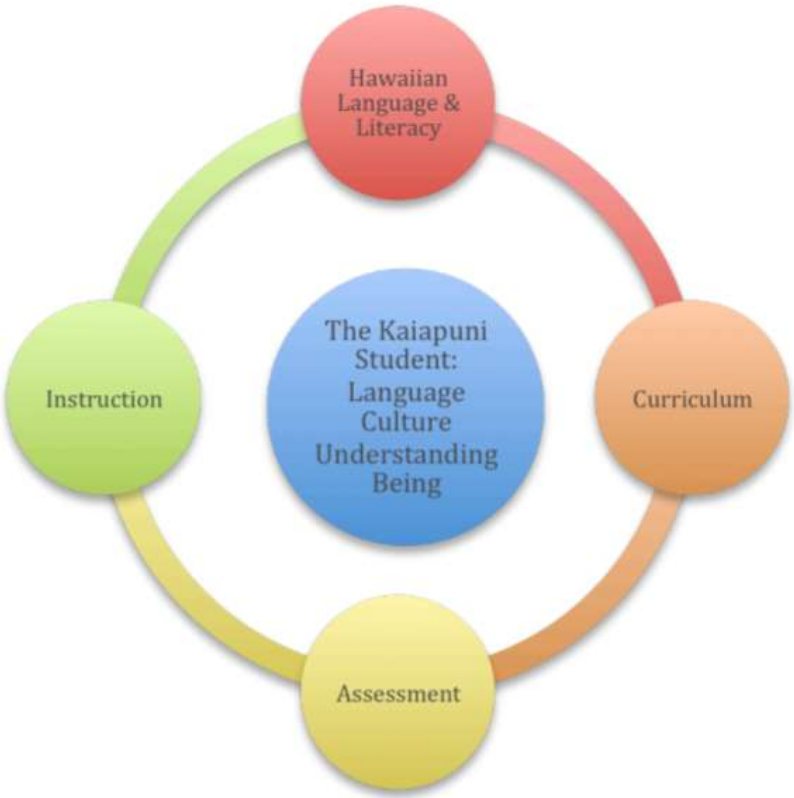


Personalization to First Language

- State assessment in native language
  - 31 states (plus DC) in math
  - 21 in science
  - 6 in SS
  - 4 in ELA
  - Almost all have Spanish versions
    - Others are Arabic, Chinese, Haitian Creole, Korean, Russian, Somali, Vietnamese
  - Direct translations or transadaptations for cultural differences

Personalization to Language and Culture

- Hawaii's Kaiapuni Assessment of Education Outcomes (KĀ'EO)
  - Grounded in language, culture, and worldview of students in Kaiapuni Hawaiian language immersion programs
  - Administered in grades 3–8 in HLA, math, and science
  - Used for community, state, and federal accountability



Adapting to Interests and Prior Knowledge

- Problem choice in the College Board essay tests (1900s)
- Continues in AP US History and European History tests
- Premise
  - Due to prior knowledge and other factors, task difficulty varies across examinees (Linn & Burton, 1994; Shavelson, Baxter, & Gao, 1993)
  - Examinees will make beneficial selections

Adapting to Interests and Prior Knowledge

- Two lines of research (1990s)
  - Choice in CR tasks (Powers & Bennett, 1999)
    - Results inconsistent
    - Some examinees don't make good choices
  - Self-adaptive MC testing (Pitkin & Vispoel, 2001)
    - Meta-analysis: choice led to marginally higher ability estimates and a small reduction in post-test anxiety

Several Paths to Adaptation

- Machine driven
  - High degree but only on a single cognitive attribute
  - Responsibility rests with test designers
  - Adding sociocultural characteristics would increase responsibility
- Assessments for specific populations
  - Many adapted versions exist
  - Disability, language, or language and culture
  - Less feasible as groups and intersectionalities increase
- Examinee driven
  - Allows cognitive and noncognitive factors to come into play
  - Shifts responsibility to the examinee

Examinee-Driven *Deep* Personalization

- One standardized assessment
  - Personalizes to cultural identity, interests, and prior knowledge
  - Uses those attributes as assets
  - Helps sustain diverse backgrounds
  - Aids instruction

Example: Deep Personalization

- Advanced Placement (AP) Art and Design Program
  - 3 courses: Drawing, 2D Art & Design, 3D Art & Design
  - Each has a portfolio examination
  - Share essential features and implications
- 3D: Students choose materials, processes, & ideas
  - May include figurative or nonfigurative sculpture, architectural models, metal work, ceramics, glasswork, installation, performance, assemblage, 3-D fabric/fiber arts, still images from videos or film, or composite images

AP 3D Art and Design Exam

- Two sections
  - Selected Works
    - Digital images of 5 artifacts showing 3D skills & synthesis of materials, processes, & ideas
    - Accompanied by written description of the ideas, materials, and processes used
  - Sustained Investigation
    - 15 digital images demonstrating the same competencies as Selected Works
    - Must show sustained investigation through practice, experimentation, and revision of materials, processes, and ideas
    - Written submission must:
      - Identify the inquiry guiding the investigation
      - Give evidence of how that inquiry directed the investigation in terms of practice, experimentation, and revision

AP 3D Art and Design

- Student negotiate choices with teachers
  - Ensure students understand portfolio requirements and assessment rubrics
- Allows choice of any artistic tradition
  - Identity aligned
  - Other

2021 AP® 3-D Art and Design Sustained Investigation (Image 4)

What defines Black hair/beauty? What are the political ties between Black hair and business? Why? How can I properly represent Black hair while still exposing the negative connotations around it? …

In this series … ,“Flexin’ My Complexion”, I am exploring the strong, historical, cultural and political tie between African American women and their hair. My work focuses on redefining Black hair in the workplace. I utilize business casual wear, specifically suit jackets, and combine them with braided hair to represent unity and equality.

2021 AP® 3-D Art and Design Sustained Investigation (Image 1)

As a transgender person, my relationship with religion is impacted by my gender identity. Through performance and installation, I sought to find connections between gendered rituals and cultural understandings of religious sanctity.

... I used fabric, water, and my body to explore rituals and gendered garments through performances with varying levels of light and texture. … In 4/8/9, mundane acts of gender presentation are framed as concerted acts of spirituality. I use low light levels and projection to emphasize the idea that gender-affirming expression can be a beacon of light in a world of darkness … Through these works, I contextualize my experiences as inherently holy, using rituals as a tool for self-empowerment and catharsis.

2021 AP® 3-D Art and Design Sustained Investigation (Image 6)

Throughout my body of work, I continuously explore: In what ways can I communicate how stress/anxiety impact me regarding my mental and physical health while having ADHD and depression?

…, my body of work creates an intimate dialogue between the wearer and the art. Each piece is physically and conceptually restrictive, communicating the feeling of suffocation and entrapment, while at the same time, incorporating the beauty and fragility found in butterflies. The movement of the butterflies emphasizes the desire to escape from unwanted feelings. Images 1-7 communicate the feeling of restriction and suffocation. … Images 11-15 start to move closer towards the ultimate goal of freedom from my anxiety.

Technical Challenges

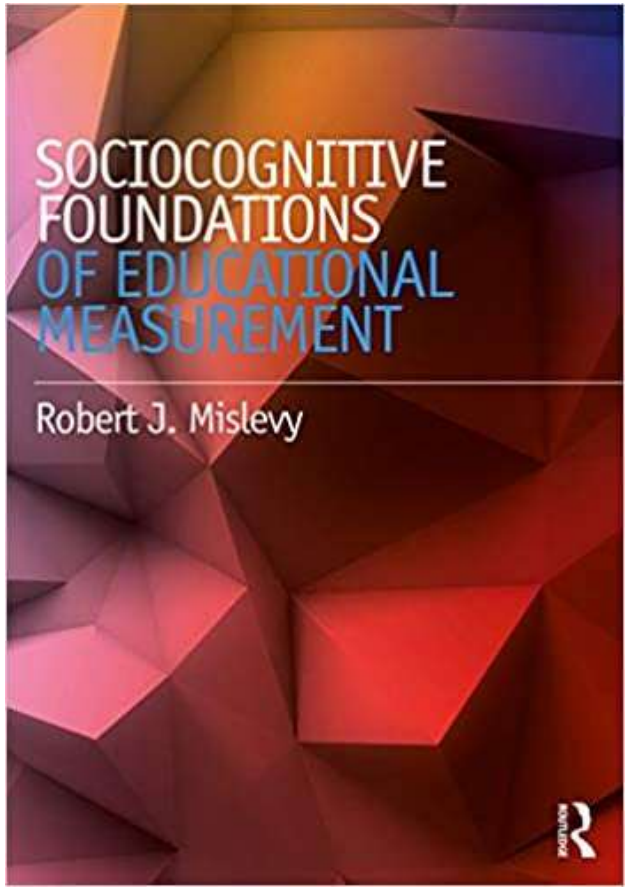
- Rater consistency, score reliability, fairness, standardization, score comparability
- Stecher (2010) literature review
  - Rater consistency achievable given:
    - Raters who understood the domain & the rating criteria
    - Task design based on clear definition of proficiency
    - Rubrics minimizing inference
    - Effective training
    - Careful monitoring
  - Score reliability depends largely on the number of tasks

3D Art and Design Portfolio Exam

- Raters are domain experts
- Rubric is general but made more concrete through extensive training with benchmark samples
- Multiple images per student judged in two collections
  - Each collection gets multiple independent ratings

Standardization

- Mislevy (2018)
  - "Conditional sense of fairness"
    - A principled basis and methodology for complex adaptation
  - Making tasks *less* comparable across individuals can produce evidence that is *more* comparable
- Sireci (2020) proposes *UNDERstandardization* to encompass:
  - Personal characteristics beyond the proficiency measured
  - Interactions of characteristics with testing conditions
  - Flexible conditions to accommodate potential interactions

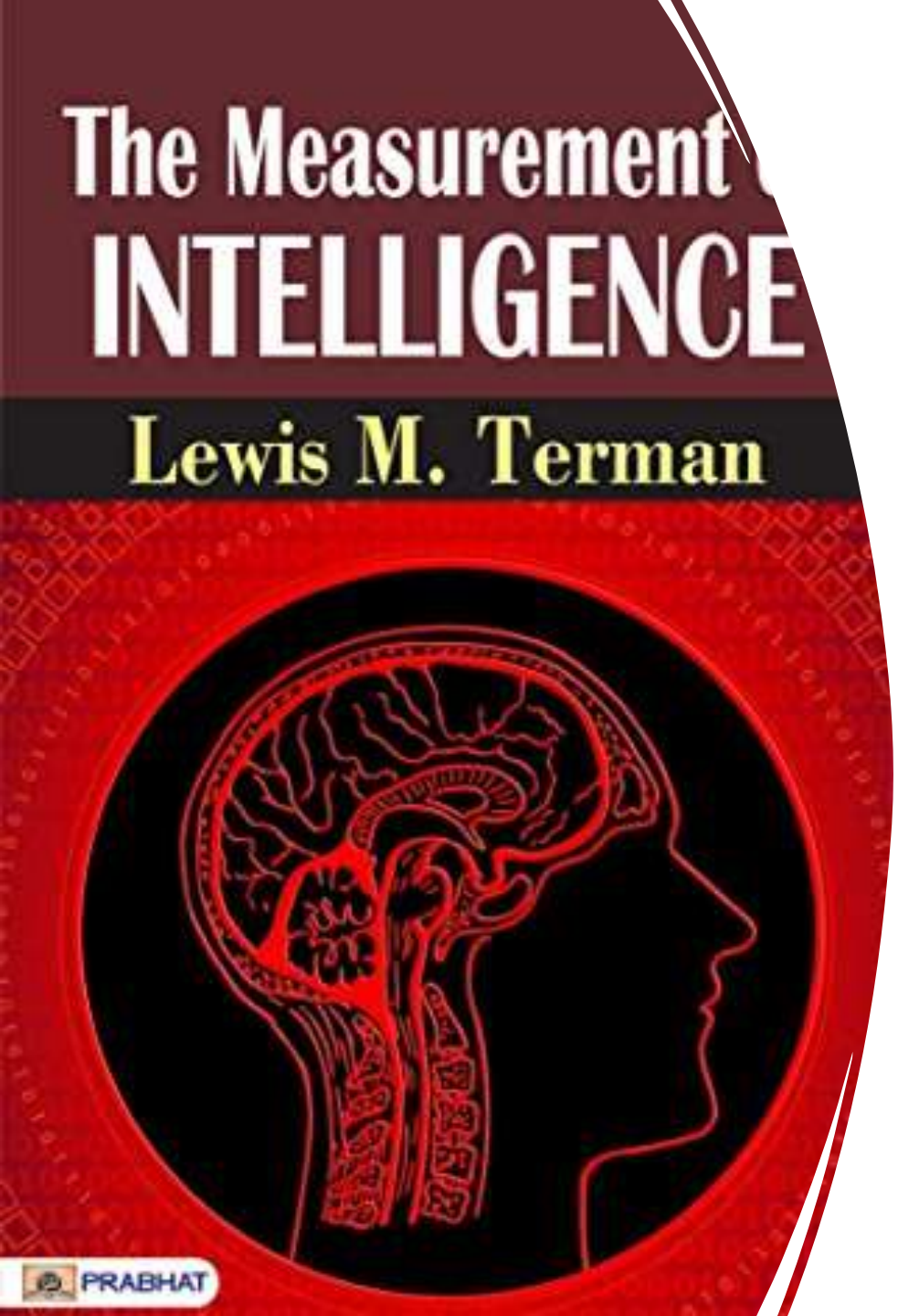


Practical Challenges

- AP Art and Design Portfolio Exams
  - Small program (~50,000 students in 2020)
  - Focused on one domain at high-school level
- AP Capstone
  - Students investigate topics in multiple disciplines, leading to a research paper and oral defense
    - Projects personalized to identity, background, and interest
    - External resources, including experts, are key
    - 31% of participants from underrepresented groups
  - 2-year course sequence
  - 2019 total volume similar Art and Design
    - Only ~1/3 participated in the 2nd course

The Role of the Teacher

- Assessments intended as learning activities
- Teachers often contribute to task creation and may participate in scoring
- Challenge
  - Teachers in high-minority schools typically less qualified
  - Sociocultural background differs markedly from students
- Can such models can be scaled in ways that allow for sociocultural responsiveness?
- PD key element of performance and portfolio programs
  - Participation an important path to improving practice (Darling-Hammond & Falk, 2013)



Principle 5: *Represent assessment results as an interaction among what the examinee brings to the assessment, the types of tasks engaged, and the conditions and context of that engagement*

- Test scores long used as indicators of personal characteristics
  - 1900s: IQ scores as genetic indicators (Brigham, 1923)
  - Trait interpretations
  - Construct interpretations
  - All presumed an underlying characteristic resident *in the person*
- Problematic
  - Personalizes achievement differences
  - Too often propagated to pernicious characterizations of groups

Sociocognitive and Sociocultural Interpretations

- Sociocognitive interpretations (Mislevy, 2018)
  - Result reflects interaction among person, history, tasks, and contexts in which they perform
  - Test performance indicates a likelihood to behave in certain ways conditional on these factors
- Sociocultural interpretations (Penuel & Shepard, 2016)
  - Add the identities, knowledge, ways of knowing, and practices valued in students' families and communities
- Both highlight that a score results:
  - *Not* from a generalized competency manifest across all tasks and situations
  - From the intersection of multiple intrinsic and extrinsic factors

Example: Represent assessment results as an interaction

- 2011 NAEP writing assessment
  - *Female students scored higher than male students at the 8th-grade level when composing online essays on demand to persuade, explain, or convey experience*
- The example:
  - Conditionalizes claim to one task type written to only three of a universe of composition purposes
  - Contextualizes claim to writing on demand and on computer
  - Avoids suggesting females generally write better than males

Pacific Standard

NEWS IN BRIEF ECONOMICS EDUCATION ENVIRONMENT SOCIAL JUSTICE FEATU

HOME › EDUCATION

SEP 20, 2018

WOMEN ARE SUPERIOR WORDSMITHS FROM AN EARLY AGE

By fourth grade, girls are better than boys at reading and, especially, writing. New research finds this gap then increases as they get older.

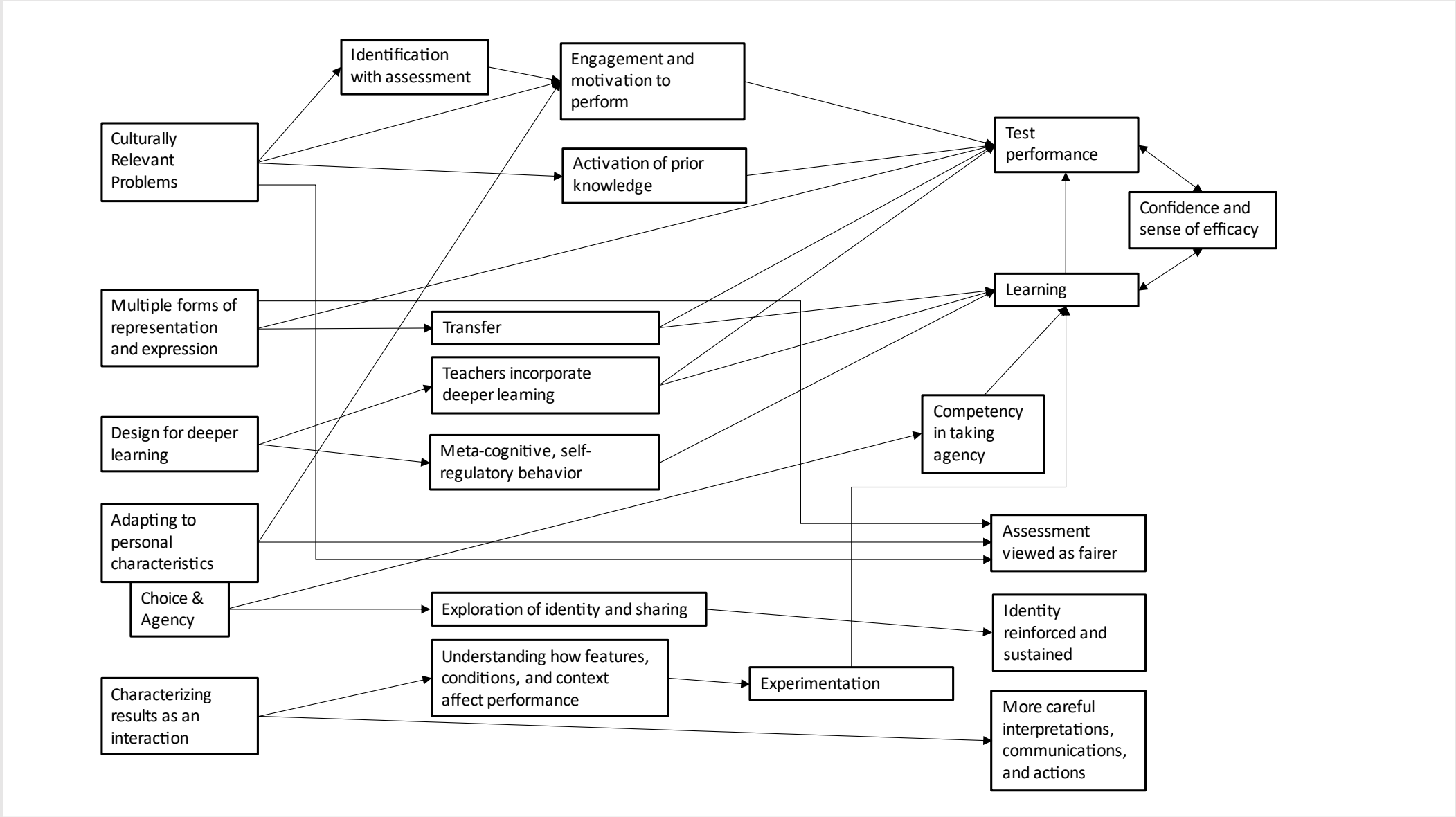
A Working Definition

Socioculturally responsive assessment:

- *includes problems that connect to the cultural identity, background, and lived experiences of all individuals, especially from traditionally underserved groups;*
- *allows forms of expression and representation in problem presentation and solution that help individuals show what they know and can do;*
- *promotes deeper learning by design;*
- *adapts to personal characteristics including cultural identity; and*
- *characterizes performance as an interaction among extrinsic and intrinsic factors.*

Socioculturally responsive assessment is assessment that people can see and affirm themselves in and from which they can learn.

An Initial Theory



A Path Forward

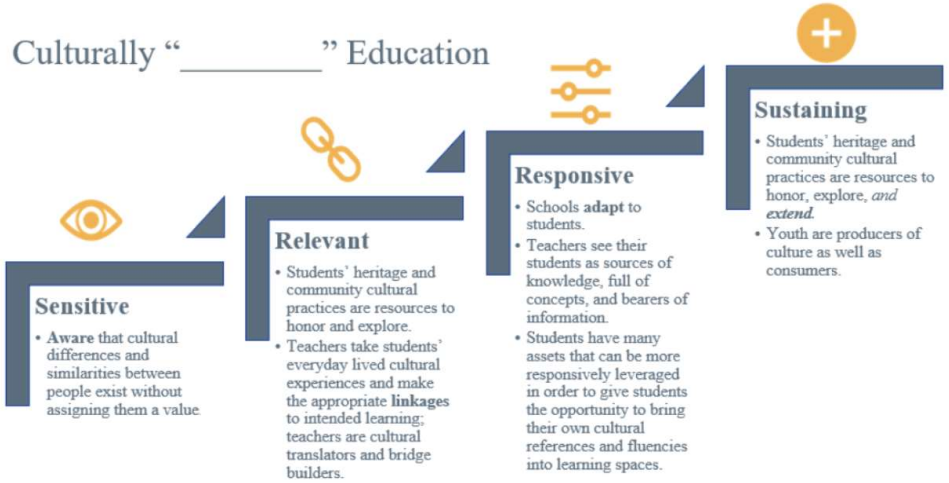
- Theory
  - Assessment design causes change in students, teachers, and others
- More likely if principles are broadly applied
  - State accountability tests
  - Admissions tests
  - Balanced assessment systems
  - Curriculum and instruction
- Principles offer basis for coherent systems
  - Multiplicity of avenues for achieving intended effects of sociocultural responsiveness

Two Fronts

- Classroom assessment
  - May allow for more complete implementations
  - Qualitative and quantitative research might find transportable features
- Testing programs
- AP Art and Design and AP Capstone are both
  - Learn from them!

Testing Programs

- Implement most achievable increments first
  - Already at *Sensitive* in Evans (2021) framework
- Incorporate socioculturally *relevant* content
  - Political
    - Content relevant to one group objectionable to another group
  - Logistical
    - Creating enough relevant content to achieve interest-group balance
    - Demands target-population involvement



A Toehold

- Every item doesn’t need to be immediately socioculturally responsive
- Start with a small number of carefully considered problems
- Don’t wait for empirical research
  - Takes years
  - Results may not be consistent
- Principles, theory, and logic may be serviceable near-term substitutes
  - Presumes reasonable consensus

A Second Toehold

- Borrow from what's well established
  - AP Art and Design
  - AP Capstone
- Operational models for deep choice
  - May not be generally feasible but modifications might be
- ELA writing assessment
  - Prompts that encourage choice of paths within a topic
  - Responses built on background, prior knowledge, interests, cultural identity
  - Supplement with choice among prompts engineered to give similar opportunities

Population-Specific Assessments

- Many disability, language, and cultural and language groups
- Proliferating test versions not logistically or economically desirable
- May be reasonable in some cases
- Precedent among international assessments
  - PISA 2018 offered in over 90 language versions

A Third Toehold

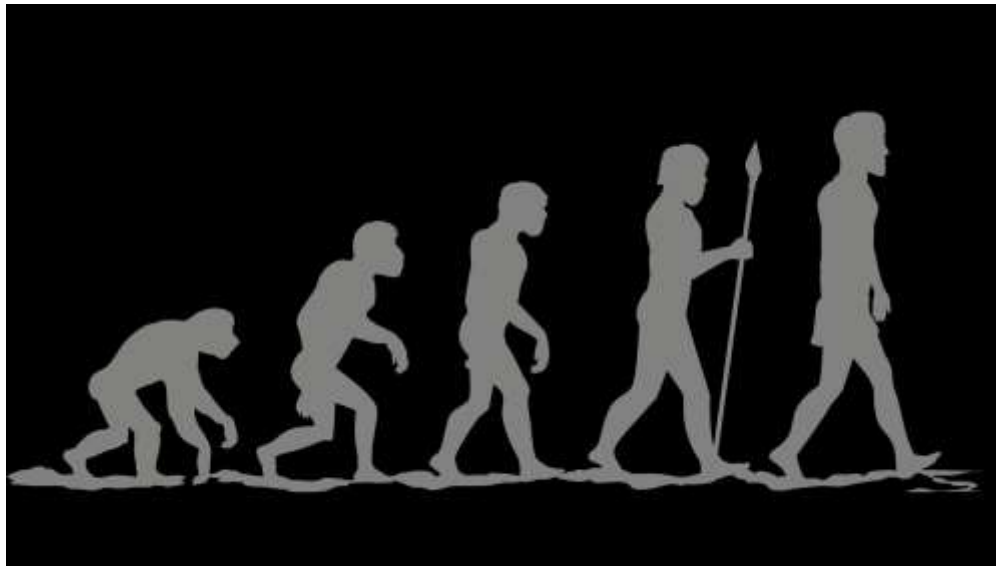
- New technology
  - Facilitates capture and submission of student work processes and products
  - Allows remote training of raters and remote rating
- Socioculturally responsive machine personalization further afield
  - Might best be pursued in the learning context
  - Port successes to testing programs

Summary

- Opposition to testing growing
- Fueled by view tests don't fit the pluralistic society US rapidly becoming
- Traditional tests not sustainable
- Socioculturally responsive assessment
  - Requires time, thought, iteration, political skill
- How do we proceed?
  - A definition
  - A theory
  - Principles for assessment design
  - Examples

Summary

- Evolve through simultaneous implementations in the classroom and in external assessments
  - Increase the cultural relevance of content
  - Provide population-specific assessment
  - Explore machine adaptation to relevant student characteristics
  - Allow learner agency
    - Problem choice
    - Deep choice



Where?

- National assessment
- State assessment
- AP
- College and graduate admissions
- K-12 classroom assessment and learning
- Teacher licensure and professional development
- Workforce

Socioculturally Responsive Assessment: What is it and What Does it Look Like?

Randy Bennett

Educational Testing Service
Princeton, NJ 08541
rbennett@ets.org

Virtual presentation at the Learning Sciences Research Institute and Department of Educational Psychology, University of Illinois Chicago, April 2022