Testing the limits of testing
• National Testing Survey: What it is, what it is for, what we have found so far
Overview

- **National Testing Survey**: What it is, what it is for, what we have found so far
- **Breadth vs depth**: Pedagogical concerns
Overview

• **National Testing Survey**: What it is, what it is for, what we have found so far

• **Breadth vs depth**: Pedagogical concerns

• **Testing testing**: Psychometric concerns
Overview

- National Testing Survey: What it is, what it is for, what we have found so far
- Breadth vs depth: Pedagogical concerns
- Testing testing: Psychometric concerns
- Moving forward: Research and redesign
Mission, instrument, & preliminary findings

Zachary Stein
Why the NTS? Why now?
Why the NTS? Why now?

- Global transformations are educational challenges
Why the NTS? Why now?
Why the NTS? Why now?

- Democracy and education
Why the NTS? Why now?
Why the NTS? Why now?

- High-stakes testing
Why the NTS? Why now?

- High-stakes testing
- Accountability dynamics
Why the NTS? Why now?

- High-stakes testing
- Accountability dynamics
- "Measurement structures the system"
Debating testing

Unproductive dichotomies
Debating testing

Unproductive dichotomies
Debating testing

Unproductive dichotomies

- sorting mechanism vs. educative aid
Debating testing

Unproductive dichotomies

- sorting mechanism vs. educative aid
- general skills vs. content knowledge
Debating testing

Unproductive dichotomies

- sorting mechanism vs. educative aid
- general skills vs. content knowledge
- work force training vs. citizenship
Beyond dichotomies

Toward dialogue
Beyond dichotomies

Toward dialogue

Beyond dichotomies

Toward dialogue

• Tests need to be built around a theory of learning.
Beyond dichotomies

Toward dialogue

- Tests need to be built around a theory of learning.
- High-stakes testing is almost never appropriate.
Beyond dichotomies

Toward dialogue

• Tests need to be built around a theory of learning.
• High-stakes testing is almost never appropriate.
• Assessment design requires research and dialogue.
Toward dialogue
• The NTS is meant to help facilitate dialogue
Toward dialogue

- The NTS is meant to help facilitate dialogue
- It can be used to monitor the impacts that assessment infrastructures have on those most affected by them
THE NATIONAL TESTING SURVEY (NTS)

We'd like you to tell us what standardized testing is doing for teachers and students in the United States. We'd also like to know what you think testing ought to do for teachers and students.

What are today's tests doing for students and teachers?

There has been a lot of discussion about the impact of testing on student learning. Some argue that objective testing is the only fair way to evaluate student learning and the quality of education being received by our students.

Others argue that current tests actually harm students by supporting instructional methods that de-emphasize essential skills — like writing, real-world problem-solving, and critical, scientific, and creative thinking.

What would you like tests to do for students and teachers?

However you feel about NCLB or standardized testing, you are likely to agree that today's tests aren't perfect. So, in addition to providing an opportunity for you to tell us what you think about today's tests, we're giving you a chance to tell us about your testing "wish list."

Take the NTS and tell us what you think, and we will share what we learn with you, and with policy makers and test developers all over the USA.

NEW NTS Launch event and workshop, May 28th & 30th, Philadelphia

Download the NTS poster
Not a public opinion poll

- It has open-ended responses.
Not a public opinion poll

- It has open-ended responses.
- It should not be used to declare the state of public opinion.
Not a public opinion poll

- It has open-ended responses.
- It should not be used to declare the state of public opinion.
- It can be used to monitor reform efforts.
Not a public opinion poll

- It has open-ended responses.
- It should not be used to declare the state of public opinion.
- It can be used to monitor reform efforts.
- It can be used as an educational tool.
• It has open-ended responses.
• It should not be used to declare the state of public opinion.
• It can be used to monitor reform efforts.
• It can be used as an educational tool.
• We will open-source all data.
What are today’s tests doing for us?

• For administrators?
• For teachers?
• For students?

[Cartoon image: A teacher in front of a blackboard that says "No child left behind," and a student asks, "Is this the test to test us for the test to see if we are ready for the test?" ]
What would ideal tests do for us?

- For administrators?
- For teachers?
- For students?
The NTS has two kinds of items.
The NTS has two kinds of items

- Survey: *Today’s tests…* (or *ideal tests…*)
The NTS has two kinds of items

- Survey: *Today’s tests… (or ideal tests…)*
  - focus on skills that students will need as adults
The NTS has two kinds of items

- Survey: *Today’s tests…* (or *ideal tests…*)
  - focus on skills that students will need as adults
  - help teachers do a better job in the classroom
The NTS has two kinds of items

- **Survey:** *Today’s tests…* (or *ideal tests…*)
  - focus on skills that students will need as adults
  - help teachers do a better job in the classroom
  - tell us how well students understand what they have learned
The NTS has two kinds of items

- **Survey:** *Today’s tests… (or ideal tests…)*
  - focus on skills that students will need as adults
  - help teachers do a better job in the classroom
  - tell us how well students understand what they have learned
- **Open-ended items**
The NTS has two kinds of items

- Survey: *Today’s tests…* (or *ideal tests…*)
  - focus on skills that students will need as adults
  - help teachers do a better job in the classroom
  - tell us how well students understand what they have learned

- Open-ended items
  - Tell a story about your experience with tests.
The NTS has two kinds of items

• Survey: *Today’s tests*… (or *ideal tests*…)
  • focus on skills that students will need as adults
  • help teachers do a better job in the classroom
  • tell us how well students understand what they have learned

• Open-ended items
  • Tell a story about your experience with tests.
  • Tell us about alternatives to current testing practices.
• 100 participants
Sample

- 100 participants
  - 68.7% female, 31.3% male
Sample

• 100 participants
  • 68.7% female, 31.3% male
  • Age: 18-80 (mean 46.29, SD 13.58)
Sample

- 100 participants
  - 68.7% female, 31.3% male
  - Age: 18-80 (mean 46.29, SD 13.58)
  - respondents are primarily teachers, parents, and researchers
Sample

- 100 participants
  - 68.7% female, 31.3% male
  - Age: 18-80 (mean 46.29, SD 13.58)
  - respondents are primarily teachers, parents, and researchers
- No differences between groups on ideal tests
Sample

- 100 participants
  - 68.7% female, 31.3% male
  - Age: 18-80 (mean 46.29, SD 13.58)
  - respondents are primarily teachers, parents, and researchers
- No differences between groups on ideal tests
- Females are a tiny bit harder on today’s tests than males: 2.43, compared to 2.82 (out of 6)
Sample

- 100 participants
  - 68.7% female, 31.3% male
  - Age: 18-80 (mean 46.29, SD 13.58)
  - respondents are primarily teachers, parents, and researchers
- No differences between groups on ideal tests
- Females are a tiny bit harder on today’s tests than males: 2.43, compared to 2.82 (out of 6)
- Alpha = .94 (ideal) and .96 (today’s)
Difference between ratings of “today’s” tests and “ideal” tests

- **strongly agree**
- **agree**
- **somewhat agree**
- **somewhat disagree**
- **disagree**
- **strongly disagree**

<table>
<thead>
<tr>
<th>Category</th>
<th>TODAY’s</th>
<th>IDEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inform administrators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>measure learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>measure relevant skills</td>
<td></td>
<td></td>
</tr>
<tr>
<td>improve learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>improve teaching</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Difference between ratings of “today’s” tests and “ideal” tests

- strongly agree
- agree
- somewhat agree
- somewhat disagree
- disagree
- strongly disagree

- ALL
- inform administrators
- measure learning
- measure relevant skills
- improve learning
- improve teaching

- TODAY's
- IDEAL
Next steps

• Launch the NTS on a large scale.
Next steps

- Launch the NTS on a large scale.
- Use in a focused context of assessment reform.
Why depth of study matters

Marc Schwartz
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2):

\[ \text{__CO(g) + __O}_2(g) \rightarrow \text{__CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1
B 1, 1, 1
C 2, 1, 2
D 1, 2, 2
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2):

$\text{2CO}(g) + \text{1O}_2(g) \rightarrow \text{2CO}_2(g)$

What coefficient values will balance the reaction shown above?

A 2, 2, 1
B 1, 1, 1
C 2, 1, 2
D 1, 2, 2
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2): 

$$2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g)$$

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

Conceptual Alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.
What are these questions assessing?

**Traditional Question**

Carbon monoxide (CO) burns readily in oxygen (O₂), forming carbon dioxide (CO₂):

\[ 2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

**Conceptual Alternative**

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.

Footnote: No significant correlation between being able to answer the traditional question correctly and the “sealed jar” problem in HS students (Agung & Schwartz, 2007).
Depth and breadth
Depth and breadth

- For most of century the argument was philosophical.
Depth and breadth

- For most of century the argument was philosophical.
- My colleagues and I conducted a study with 8341 college freshman
Depth and breadth

- For most of century the argument was philosophical.
- My colleagues and I conducted a study with 8341 college freshman
  - attending one of 55 randomly chosen US colleges or universities;
Depth and breadth

- For most of century the argument was philosophical.
- My colleagues and I conducted a study with 8341 college freshman
  - attending one of 55 randomly chosen US colleges or universities;
  - enrolled one of three introductory science courses (biology, chemistry or physics), and
Depth and breadth

• For most of century the argument was philosophical.
• My colleagues and I conducted a study with 8341 college freshman
  • attending one of 55 randomly chosen US colleges or universities;
  • enrolled one of three introductory science courses (biology, chemistry or physics), and
  • who self-reported on over 40 items regarding their experience in high school science
Variables
Variables

- We focused on two variables:
• We focused on two variables:
  • **Independent variable of interest**: Amount of time they spent on various topics in their HS science class.
Variables

- We focused on two variables:
  - **Independent variable of interest**: Amount of time they spent on various topics in their HS science class.
  - **Dependent variable (i.e., outcome measure)**: Student’s final grade in the introductory college course (as noted by professor)
Variables

- We focused on two variables:
  - **Independent variable of interest:** Amount of time they spent on various topics in their HS science class.
  - **Dependent variable (i.e., outcome measure):** Student’s final grade in the introductory college course (as noted by professor)
- What was the outcome?
Does it matter? Depth or breadth: Does it matter?

Difference from average college grade in science
Breadth present
Depth present
Biology                 Chemistry                  Physics
3
2
1
0
-1
-2
-3

Schwartz, Sadler, Sonnert, & Tai (2009)
Does pedagogy matter?

Schwartz, Sadler, Sonnert, & Tai (2009)

Difference in years of HS science study

% of high school students
Does pedagogy matter?

Depth present, breadth absent has roughly the same impact as taking an additional semester of study in the same subject.

Schwartz, Sadler, Sonnert, & Tai (2009)
Does pedagogy matter?

**Depth absent, breadth present**
Has roughly the same impact as ending the school year in March.

**Depth present, breadth absent**
has roughly the same impact as taking an additional semester of study in the same subject

Schwartz, Sadler, Sonnert, & Tai (2009)
Testing testing

Theo L. Dawson
Three reasons for concern
Three reasons for concern

- Reliability
- Construct validity
- Objectivity
Three claims
Three claims

- **Reliability**: Today’s tests are the most reliable ever made.
Three claims

- **Reliability**: Today’s tests are the most reliable ever made.
- **Construct validity**: Today’s tests measure what they set out to measure.
Three claims

- **Reliability**: Today’s tests are the most reliable ever made.
- **Construct validity**: Today’s tests measure what they set out to measure.
- **Objectivity**: Today’s tests treat everyone the same.
Reliability

How confident can we be that one student’s score on a test is really higher or lower than another student’s score?

(Reliability is expressed as a decimal between 0 and 1)
Reliability

How confident can we be that one student’s score on a test is really higher or lower than another student’s score?

(Reliability is expressed as a decimal between 0 and 1)

- A reliability of .85 allows us to distinguish up to 3 different levels of performance.
- A reliability of .90 allows us to distinguish up to 6 different levels of performance.
- Our most reliable tests are in the .90 – .95 range.
- That’s all reliability can tell us.
Construct validity

Does this test measure what it is intended to measure?
Construct validity

Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
Construct validity

Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
- Do its items cover the appropriate range of difficulty?
Construct validity

Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
  - Do its items cover the appropriate range of difficulty?
  - Do the items require knowledge or skills that are not relevant to the targeted knowledge area?
Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
  - Do its items cover the appropriate range of difficulty?
  - Do the items require knowledge or skills that are not relevant to the targeted knowledge area?
    - writing skills on a math test
Construct validity

Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
- Do its items cover the appropriate range of difficulty?
- Do the items require knowledge or skills that are not relevant to the targeted knowledge area?
  - writing skills on a math test
  - items that assume particular life experiences (bias)
Construct validity

Does this test measure what it is intended to measure?

- Do its items do a good job covering the content or skills of the knowledge area?
  - Do its items cover the appropriate range of difficulty?
  - Do the items require knowledge or skills that are not relevant to the targeted knowledge area?
    - writing skills on a math test
    - items that assume particular life experiences (bias)
    - of particular concern in high-stakes testing
Does this test treat *everyone* the same?
Objectivity

Does this test treat everyone the same?

- Does every student get the same chance to show off what he or she has learned?
Does this test treat *everyone* the same?

- Does every student get the same chance to show off what he or she has learned?
- Is the scoring process free of bias?
Objectivity

Does this test treat *everyone* the same?

- Does every student get the same chance to show off what he or she has learned?
- Is the scoring process free of bias?
- Does every student take the test under the same conditions?
Three reasons for concern

Why we should be concerned about our reliance on reliability, construct validity, and objectivity.
Three reasons for concern

Why we should be concerned about our reliance on reliability, construct validity, and objectivity.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
Why we should be concerned about our reliance on reliability, construct validity, and objectivity.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
- Tests focus on a restricted range of skills and knowledge areas.
Three reasons for concern

Why we should be concerned about our reliance on reliability, construct validity, and objectivity.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
- Tests focus on a restricted range of skills and knowledge areas.
  - Concerns about construct validity (especially as it relates to bias) have led to narrow representations of skills in some knowledge domains (like mathematics and science).
Three reasons for concern

Why we should be concerned about our reliance on *reliability*, *construct validity*, and *objectivity*.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
- Tests focus on a restricted range of skills and knowledge areas.
  - Concerns about construct validity (especially as it relates to bias) have led to narrow representations of skills in some knowledge domains (like mathematics and science).
  - Concerns about objectivity have increased reliance upon the multiple choice item type, which is most useful for assessing:
Why we should be concerned about our reliance on reliability, construct validity, and objectivity.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
- Tests focus on a restricted range of skills and knowledge areas.
  - Concerns about construct validity (especially as it relates to bias) have led to narrow representations of skills in some knowledge domains (like mathematics and science).
  - Concerns about objectivity have increased reliance upon the multiple choice item type, which is most useful for assessing:
    - memory for factual knowledge, and
Three reasons for concern

Why we should be concerned about our reliance on reliability, construct validity, and objectivity.

- Reliabilities are only high enough to distinguish 3 to 6 levels.
- Tests focus on a restricted range of skills and knowledge areas.
  - Concerns about construct validity (especially as it relates to bias) have led to narrow representations of skills in some knowledge domains (like mathematics and science).
  - Concerns about objectivity have increased reliance upon the multiple choice item type, which is most useful for assessing:
    - memory for factual knowledge, and
    - the ability to accurately apply rules/algorithms/formulas.
Ideal tests

- Are reliable and assess what they claim to assess.
Ideal tests

- Are reliable and assesses what they claim to assess.
- Balance concerns about objectivity with concerns about relevance.
Ideal tests

- Are reliable and assess what they claim to assess.
- Balance concerns about objectivity with concerns about relevance.
- Are relevant, because they…
ideal tests

- Are reliable and assesses what they claim to assess.
- Balance concerns about objectivity with concerns about relevance.
- Are relevant, because they...
  - are part of the conversation between students and teachers that tells teachers what students are most likely to benefit from learning next, and
Ideal tests

- Are reliable and assess what they claim to assess.
- Balance concerns about objectivity with concerns about relevance.
- Are relevant, because they...
  - are part of the conversation between students and teachers that tells teachers what students are most likely to benefit from learning next, and
  - assess skills and ideas that support future learning and success in adult life.
Ideal tests

- Are reliable and assess what they claim to assess.
- Balance concerns about objectivity with concerns about relevance.
- Are relevant, because they...
  - are part of the conversation between students and teachers that tells teachers what students are most likely to benefit from learning next, and
  - assess skills and ideas that support future learning and success in adult life.
- Know their place.
To be relevant, tests must
To be relevant, tests must focus on skills and knowledge that prepare students for adult life (including skills for life-long learning),
To be relevant, tests must

- focus on skills and knowledge that prepare students for adult life (including skills for life-long learning),
- help teachers teach by providing them with useful information about their students, and
Relevance

• To be relevant, tests must
  • focus on skills and knowledge that prepare students for adult life (including skills for life-long learning),
  • help teachers teach by providing them with useful information about their students, and
  • help students learn by pointing out strengths, weaknesses, and “what comes next”.
Traditional Question
Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2):

\[ 2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

Conceptual Alternative
A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.
Traditional Question
Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2):

\[ 2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

Has the student memorized this fact?

Conceptual Alternative
A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.
**Traditional Question**

Carbon monoxide (CO) burns readily in oxygen (O₂), forming carbon dioxide (CO₂):

\[ 2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

---

**Conceptual Alternative**

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.
**Traditional Question**

Carbon monoxide (CO) burns readily in oxygen (O2), forming carbon dioxide (CO2):

\[ 2\text{CO}(g) + 1\text{O}_2(g) \rightarrow 2\text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

---

**Conceptual Alternative**

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.

---

Has the student memorized this fact?  
Has the student understood these concepts well enough to solve a problem?
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O₂), forming carbon dioxide (CO₂):

\[
\text{CO}(g) + \text{O}_2(g) \rightarrow \text{CO}_2(g)
\]

What coefficients will balance the reaction shown above?

A. 2, 2, 1
B. 1, 1, 1
C. 2, 1, 2
D. 1, 2, 2

Conceptual Alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.
b. The pan will not move.
c. Move down.
d. First move up and then down.
e. There is not enough information to answer the question.

Open-ended alternative

"The reaction makes a gas that is lighter than the vinegar or the soda, so the right pan will go up a little."
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O₂), forming carbon dioxide (CO₂):

\[ \text{CO(g)} + \text{O}_2(g) \rightarrow \text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

Conceptual Alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.

Open-ended alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

“The reaction makes a gas that is lighter than the vinegar or the soda, so the right pan will go up a little.”

Which concepts is this student working with?

How does she understand these concepts?

What is her line of reasoning?

How well does she explain her thinking?
Traditional Question

Carbon monoxide (CO) burns readily in oxygen (O₂), forming carbon dioxide (CO₂):

\[ \text{CO}(g) + \text{O}_2(g) \rightarrow \text{CO}_2(g) \]

What coefficient values will balance the reaction shown above?

A 2, 2, 1  
B 1, 1, 1  
C 2, 1, 2  
D 1, 2, 2

Conceptual Alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

a. Move up.  
b. The pan will not move.  
c. Move down.  
d. First move up and then down.  
e. There is not enough information to answer the question.

Open-ended alternative

A scale is balanced with two sealed jars. The left pan has a sealed jar containing vinegar and 5 grams of baking soda is lying outside. The right pan has a sealed jar containing vinegar and the same amount of baking soda inside the jar. As the baking soda fizzes, what will happen to the pan with the fizzing baking soda?

“The reaction makes a gas that is lighter than the vinegar or the soda, so the right pan will go up a little.”

Which concepts is this student working with?

How does she understand these concepts?

What is her line of reasoning?

How well does she explain her thinking?
Knowing their place
Knowing their place

- Given that…
Knowing their place

• Given that…
  • all tests must focus on specific sets of skills or knowledge (construct validity) that represent a limited range of capabilities,
Knowing their place

• Given that…
  • all tests must focus on specific sets of skills or knowledge (construct validity) that represent a limited range of capabilities,
  • all tests are limited in their ability to distinguish between levels of performance (reliability), and
Knowing their place

• Given that...

  • all tests must focus on specific sets of skills or knowledge (construct validity) that represent a limited range of capabilities,
  • all tests are limited in their ability to distinguish between levels of performance (reliability), and
  • all tests are further constrained in what they can measure by the way which they are administered and scored,
Knowing their place

- Given that...
  - all tests must focus on specific sets of skills or knowledge (construct validity) that represent a limited range of capabilities,
  - all tests are limited in their ability to distinguish between levels of performance (reliability), and
  - all tests are further constrained in what they can measure by the way which they are administered and scored,
- It is unethical to use any single test, single type of test, or tests on their own as the sole criteria for making any high stakes decision.
Ecological validity
National Testing Survey

Evidence-based testing

Zachary Stein
• a canary in the mine?
Calls for a radical redesign
Calls for a radical redesign

- **From the Left**: Revitalizing democracy and increasing educational equity and opportunity
Calls for a radical redesign

- **From the Left:** Revitalizing democracy and increasing educational equity and opportunity
- **From the Right:** Global economic competitiveness and the knowledge economy
Calls for a radical redesign

- **From the Left:** Revitalizing democracy and increasing educational equity and opportunity
- **From the Right:** Global economic competitiveness and the knowledge economy
- **From the center of the Academy:** National Research Council Committee on the Foundations of Assessment
Priority 1

• Re-fashion the accountability dynamics of the current infrastructure
Priority 1

• Re-fashion the accountability dynamics of the current infrastructure
  • Respect teachers as professionals
Priority 1

- Re-fashion the accountability dynamics of the current infrastructure
  - Respect teachers as professionals
  - Respect students as learners
Priority 2

- Fund research and development into alternative infrastructures
• Fund research and development into alternative infrastructures
  
  • MBE research school networks
Priority 2

- Fund research and development into alternative infrastructures
  - MBE research school networks
  - basic research
Priority 2

- Fund research and development into alternative infrastructures
  - MBE research school networks
  - basic research
  - ecological validity
Priority 2

• Fund research and development into alternative infrastructures
  • MBE research school networks
  • basic research
  • ecological validity
  • experimental pilot programs
Priority 2

- Fund research and development into alternative infrastructures
  - MBE research school networks
  - basic research
  - ecological validity
  - experimental pilot programs
  - the NTS as one index of impact
Priority 3
Priority 3

- Facilitate discourse in the public sphere about what is possible and preferable for our educational system
Priority 3

- Facilitate discourse in the public sphere about what is possible and preferable for our educational system
  - find ways to bring everyone to the table and change the nature of the debate
Priority 3

- Facilitate discourse in the public sphere about what is possible and preferable for our educational system
  - find ways to bring everyone to the table and change the nature of the debate
  - use the NTS as one catalyst
Priority 4
• Re-vision the nature of schooling
Priority 4

- Re-vision the nature of schooling
  - industrial dinosaurs in the post-industrial information age