Assessing Reading Comprehension and Vocabulary:

What does the research tell us?
What should we do in our schools?

P. David Pearson
UC Berkeley
Given at the U of MN on Nov. 17, 2005
A very short history lesson
What, if any, are the “research-based” findings on RC and Voc assessment?
What DO we do in the name of comprehension and vocabulary assessment?
What research needs to be conducted in the next 5 years?
What should a school or a district do while we wait for gold standard assessments to be developed, validated, and enacted?
Why now?

- Renewed interest among scholars
  - Rand report
  - National Reading Panel
  - Reading First
- Uneasiness among practitioners that the code, as important as it is, may not be the point of reading
- Comprehension is the most important outcome of reform
- Vocabulary rears its head for kids who struggle
National thirst for accountability requires impeccable measures (both conceptually and psychometric)

✓ When the stakes are high, so too must be the standards

Pleas of teachers desperate for useful tools (need a tool that does for RC what running records and fluency assessments have done for word level processes)*

*don’t hear this for vocabulary
While we definitely need better theoretically motivated measures of comprehension and vocabulary,…

We desperately need the school/classroom tool.

A measure that serves a diagnostic or monitoring function may be more critical than a conceptually elegant outcome measure.
We want to access the thing itself, the “click”

But

We only ever see its residue, its wake, its artifacts

We are stuck with artifacts

✓ Require them to tell us whether they understood
✓ Require them to tell us what they understood or remember
✓ Quiz them on the details
✓ Request the big ideas
Most RC measures interpose some other skill or capacity between the act and the evidence

- Writing
- Talking
- Using (as in an application task)
- The conventions of multiple-choice assessments (they may provide excessive scaffolding)
- These interposed processes inevitably compromise our capacity to draw inferences about comprehension (as the ineffable thing itself), either as a generic and a passage specific enterprise
What would it mean to meet the gold standard in assessment research?

Unlike instruction, we are NOT looking at randomized field trials.

Instead, the gold standard for an assessment is meeting the **construct** validity test.
We show that our test of RC is consistent with what our theory predicts about relationships among various hypothesized components of and precursors to reading comprehension.

For example: readers do not recall specific details about an idea unless they also recall and name the idea (Rumelhart, 1977).

For example: readers do not answer a question about a specific part of the text unless they first demonstrate accurate decoding of that text segment.

For example: Kintschian model: a reader ought to have built a text base as a prerequisite for a situation model for a text.
The number of word meanings one can access for a given text, the greater the overall comprehension of that text.

Vocabulary depth:

This:
A gendarme is a kind of
  • Person
  • Plant
  • Machine

Before this:
A gendarme is a kind of
  • Policeman
  • Fireman
  • Mayor
Weak version

When we look across all the evidence we have (face validity, concurrent validity, predictive validity, internal factor and cluster analyses, common sense), things seem, on average, to point to this version of our theory and therefore this set of sub-skill assessments.
We have yet to get the strong version of either

We do have some candidate versions of the weak version…

- An obscure but elaborate set of analyses of relationships among reading performance variables over time (Meyer, Linn, & Hastings, circa 1988 at CSR)
- A lot of the older factor analytic studies (later…)
What all this means is…

When you leave here today, you should

1. acknowledge and live with the weak state of our knowledge and certainty about the validity of our measures of reading comprehension and vocabulary

2. work to create the kinds of tests our teachers and kids deserve.

3. be prepared to make some strategic choices about what you do in your district or school
Meanwhile, back at the LAST turn of the century....
Conclusion: Any approach to comprehension assessment you might conjure up, even in your most enlightened moments, has a precedent that is at least 75 years old.

Novelty is a conceit but not a virtue.
Check two of the following statements with the same meaning as the quotation above.

- To know right is to do the right.
- Our speculative opinions determine our actions.
- Our deeds often fall short of the actions we approve.
- Our ideas are in advance of our everyday behavior.

Note the multiple correct answers.

From Thurstone, undated circa 1910
1916 Kansas Silent Reading Test*

- “fill in the blanks”
- some verbal logic problems
  - If A is X and B is Y, what will…
- some procedural tasks
  - Use your pencil to draw a line between X and Y
- Complete as many tasks below as possible in a limited 7 minutes.

*The first published standardized comprehension test.

Note the fluency (speeded) character
1917: Thorndike

- **Reading as Reasoning**
- Basically an error analysis leading to a set of categories and a theory
- Understanding a paragraph is like solving a problem in mathematics. It consists in selecting the right elements in the situation and putting them together in the right relations, and also with the right amount of weight or influence or force of each
- Key findings: overpotency and underpotency of words
(a) failure to understand the question
(b) failure to isolate elements of “an involved statement” read in context
(c) failure to associate related elements in a context
(d) failure to grasp and retain ideas essential to understanding concepts
(e) failure to see setting of the context as a whole
(f) other irrelevant answers
A panoply of measures

- Courtis (1914): proportion of all words in the text remembered*
- Chapman (1924): First example of error detection:
  - Find the statements in part 2 that do not fit the statements in part 1 of the paragraph.

*Not unlike DIBELS today
Enter Psychometrics in the late 1930s

1935: IBM introduced the IBM 805 scanner
✔ Cemented multiple-choice format
✔ Changed the SAT forever

1935: Kelley: Factor Analysis

1944: Davis: Fundamental Factors in Reading Comprehension
# Davis 1944

<table>
<thead>
<tr>
<th>Word meanings</th>
<th>Text based questions with paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word meanings in context</td>
<td>Draw inferences about content</td>
</tr>
<tr>
<td>Follow passage organization</td>
<td>Literary devices</td>
</tr>
<tr>
<td>Main thought</td>
<td>Author’s purpose</td>
</tr>
<tr>
<td>Answer specific text-based questions</td>
<td></td>
</tr>
</tbody>
</table>

Word factor and a reasoning factor
Other Factor Analyses

- Harris 1948: found a single factor
- Derrik (1953) found 3
- Holmes (1950s) sub-strata factor theory
- Hunt (1957) Vocabulary was everything
- Schreiner, Hieronymus, and Forsyth (1971): No differentiation among paragraph meaning, cause and effect, reading for inferences, and selecting main ideas BUT separate LC and lower level processing
- Davis (1968, 1972)

Dominant finding (word factor, gist factor, reasoning factor)
Wilson Taylor (1953): every 5th word
✓ More importantly, it was an attempt to remove human judgment from the assessment process.
✓ Pick a starting point in the text, let the randomization process do its work
✓ Doesn’t matter where you start

Bormuth (1966): the basis of readability research: average word length and average sentence length best predicted cloze fill in rates*

*still with us in lexile scaling
Modifications to Cloze

- Allow synonyms to serve as correct answers
- Delete only every 5th content word (leaving function words intact)
- Use an alternative to every 5th word deletion
- MAZE: MC for the blanks
- Macro cloze: phrases
- Delete words at the end of sentences or paragraphs and provide a set of choices from which examinees are to pick the best answer
The conceptual death of cloze

Shanahan, Kamil, & Tobin (1983): not sensitive to “intersentential” comprehension

No differences when sentences were scrambled within or across passages or presented in isolation

So…how could it measure comprehension, which, on the face of it, requires reasoning across sentences?
Despite strong evidence showing its invalidity, it still survives

- DRP
- Stanford Diagnostic
- Lots of other individual and group tests
- Strong in ESL assessment

Why?
- Feels right, feels good
- Simplicity of scoring and interpretation
Passage Dependency

P_{passage} - P_{isolation}

A quiet stir in the late 60s and early 70s

The basic idea is that if you read the passage, you ought to get the item right (even if an inference) more often than if you don’t read the passage.

Died in the wake of Schema Theory’s embrace of prior knowledge--which encouraged us to embrace, not lament, the PK-Comprehension relationship.
Criterion-referenced assessment

- Make a virtue out of sub-skills
- Took the notions of mastery learning coming out of Carroll, Gagné and Bloom
- Define sets of subskills
- Set a level of mastery
- Test-teach-test
- Assumes a componential skill view of reading
- Data: Bloom’s experiments with Ed Psy courses
The children wanted to make a book for their teacher. One girl brought a camera to school. She took a picture of each person in the class. Then they wrote their names under the pictures. One boy tied all the pages together. Then the children gave the book to their teacher.

1. What happened first?
   a. The children wrote their names
   b. Someone brought a camera to school
   c. The children gave a book to their teacher

2. What happened after the children wrote their names?
   a. A boy put the pages together.
   b. The children taped their pictures.
   c. A girl took pictures of each person

3. What happened last?
   a. The children wrote their names under the pictures.
   b. A girl took pictures of everyone.
   c. The children gave the book to their teacher.

This became the “bread and butter” of basal assessments
Reactions to this movement

- Provided fuel for the constructivist reforms that were already gathering momentum
- Died in the early 90s basals for about 6 years
- Only to be revived recently in the name of standards-based assessments
The Cognitive Revolution

- The powerful impact of schema
- The evolution of text analytic systems
  - Story grammars ala Stein & Glenn
  - Propositional analysis of texts ala Kintsch & vanDijk
- Inference taxonomies ala Trabasso
The Impact of Cognitive Science on Assessment

- more attention to the role of prior knowledge
- attention to text structure (in the form of story maps and visual displays to capture the organizational structure of text)
- the introduction of metacognitive monitoring
- used to critique the existing assessment traditions on the way to new assessments
Cognitive perspectives claim that we had

- Paid too much attention to measurement theory and
- Not enough to reading theory
Select, not construct, texts for understanding
Can’t tinker with the text to rationalize items and distractors
More than one right answer

How does Ronnie reveal his interest in Anne?

✓ Ronnie cannot decide whether to join in the conversation.
✓ Ronnie gives Anne his treasure, the green ribbon.
✓ Ronnie gives Anne his soda.
✓ Ronnie invites Anne to play baseball.
✓ During the game, he catches a glimpse of the green ribbon in her hand.
How does Ronnie reveal his interest in Anne?

- (2)(1)(0) Ronnie cannot decide whether to join in the conversation.
- (2)(1)(0) Ronnie gives Anne his treasure, the green ribbon.
- (2)(1)(0) Ronnie gives Anne his soda.
- (2)(1)(0) Ronnie invites Anne to play baseball.
- (2)(1)(0) During the game, he catches a glimpse of the green ribbon in her hand.
Include

- Metacognition
- Habits, attitudes, and dispositions
Some findings

Comprehension plus PK, Metacognition, Habits/Attitude

Factor Analyses (Pearson, et al, 1991) demonstrated three reliably separable factors

- Metacognitive stances
- habits/attitudes items
- a combination of the comprehension and prior knowledge items (could not separate them)
Went the way of all tests that challenge the conventional wisdom
No one got the more than one right answer metaphor
Intentionally validated for group decisions not individual (as accountability changed…)
Not good to teach to (e.g. metacognitive items)
Learning and understanding are inherently social.
Assessment should be responsive, interactive, and dynamic.
Texts are inherently political documents with points of view and agendas and authors.
Rosenblatt: Reader, text, and poem.
Langer: Into, through, and beyond.
If you were explaining what this essay is about to a person who had not read it, what would you say?

What do you think is important or significant about it?

What questions do you have about it?

This is your chance to write any other observations, questions, appreciations, and criticisms of the story’’
The demise of performance assessment in wide-scale

- The social aspect: Whose work is it anyway?
- Generalizability: Too passage specific
- Expense: Scoring and rubric development
- Invasion of privacy (don’t ask my kid about his inner thoughts)

- The legacy:
  - Mixed models
  - Classroom assessment
NAEP

Circa 1970 (my first encounter with NAEP was a talk at a PDK meeting on this campus by Jack Merwin)

Goal free evaluation

What you see is what you get

Report the p-values of individual items and let the readers conclude what they will
Demonstrate the ability to show comprehension of what was read

✓ analyze what is read, use what is read
✓ reason logically
✓ make judgments
✓ have attitude/interest in reading
value reading and literature
comprehend written works
respond to written works in interpretive and evaluative ways
apply study skills
FORMING INITIAL UNDERSTANDING
✓ Which of the following is the best statement of the theme of the story

DEVELOPING INTERPRETATIONS
✓ What caused this event

PERSONAL REACTION AND RESPONSE
✓ How did this character change your ideas of _____

READER TEXT CONNECTIONS

DEMONSTRATE CRITICAL STANCE
✓ What could be added to improve the author’s argument

Note: Vocabulary has always been embedded in initial understanding or critical stance.
The 1990s framework does not pass psychometric muster (no structural independence of the stances)

Not much information at the lower end of the performance scale (no floor)

Item format: Do CR items add any value over MC to the information gained?

Not if they are MC in disguise?
# Brand New NAEP Framework

<table>
<thead>
<tr>
<th>Aspects ➔</th>
<th>General Understanding</th>
<th>Interpretation</th>
<th>Reader Text Connections</th>
<th>Examining content &amp; structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contexts  ➖</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For Literary Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To perform a task</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
New Initiatives

- Lots of psychometric work
- Lots of conceptual work
- Share a few examples
The standards for good assessment, especially those dealing with instructional sensitivity, are critical.

✓ Notice that in most of our work, we assume the validity of our measures and test the validity of the interventions.

✓ What if we turned that around?
Starting Over

Why?

- Our current collection of assessments are atheoretical
- They do not map onto any credible theory of the reading comprehension process

Driven by
- Tradition (a by product of concurrent validity)
- Convenience (it’s there)
- Efficiency (it’s quick and dirty)
Starting over

Go back to a set of theoretical conceptualizations of comprehension
  ✓ Component Skill Models
  ✓ Construction-Integration models
  ✓ Executive Control Models
  ✓ Sociocultural Models

Convene a Blue Ribbon Panel to mine each for assessment implications

Apply each set of implications to a common set of passages to create a set of alternative theory-based assessments

Examine internal covariation and external validity.
Key step: Develop a “gold standard” for comprehension

How do we get as close as possible to that ineffable phenomenon—the click of comprehension?

My gold standard candidate: Some on-line assessment of both the content (ideas in text) and the affect (phenomenological sense) of comprehension (akin to the write alongs)

✓ So what’s new in this section that you didn’t know before…?

✓ So on a scale from 1-5, how would you rate your grasp of the ideas in this section
More steps in validation

- Examine the concurrent validity of the assessment models generated from each theoretical perspective in relation to the gold standard.
- Develop a grand theory to test.
- Conduct a full-scale, theory-based construct validation.
- Be open to the possibility of a mixed model.
Conclusion leading to today’s situation

- We have traveled far, sometimes on new roads and sometimes on old.
- Virtually all the old forms of assessment survive, even flourish because of their
  - Psychometric properties
  - Efficiencies
- And because challengers often fail to meet either psychometric or efficiency standards
Conclusion about research

We seem poised to re-energize ourselves in this important enterprise.

To build assessments that can meet the most rigorous of both measurement and conceptual standards.

A welcome challenge.
So what should a school or district do while we wait for the millennium of comprehension assessment

We cannot invoke the strong version of construct validity because we don’t have a single measure that can meet it.

We could invoke the criterion validity standard, but that just perpetuates some version of the status quo. We don’t know have a gold standard to decide among pretenders to the throne.
Here are some standards we could invoke even now…

- **Reliability**
- **Multiple indicators of criterion validity** (concurrent and predictive)
- **Instructional sensitivity**
  - If I teach comprehension well (using the well-validated methods you will learn about today and tomorrow), will the measure show the growth that is occurring?
- **Consequential validity**
  - If I use the test to categorize kids, diagnose and prescribe instruction, or monitor progress along the way, will students get the instruction they need and deserve?
Were there but world enough and time...

- [ ] Yes
- [ ] No
So what is a body to do?

- The Woodworth, MI system
- Benchmark assessment, used 2 to 3 times per year
- Scored in PD sessions, across classrooms and across grades
- Create a school culture
School-wide Comprehension Assessment

- Instructionally embedded
- Multiple choice questions
  - Individual texts
  - Cross texts
- Written Response to Reading
  - Position taken in response to the prompt question
  - Support from personal experience
  - Support from texts
Listening: Sister Anne’s Hands

Sister Anne’s Hands
MARYBETH LORBIECKI • illustrated by K. WENDY POPP
This story is mostly about…
Sister Anne showed determination when she said…
What did Sister Anne mean when she said, “For me, I’d rather open my door enough to let everyone in”?
The children learned much from Sister Anne. This selection tells us that…
Kate Shelly and the Midnight Express
An important lesson of this story is…
How are Kate and her mother different?
In this selection, how do you know Kate showed determination and bravery when crossing the Des Moines River Bridge?
Because Kate followed through, how would you predict she will face problems in the future?
What dialogue does the author use to show you Kate has determination?
How do you know this story takes place in the past?
A Day’s Work

by Eve Bunting

Illustrated by Ronald Himler
Multiple Choice Question Stems

- By showing determination, Francisco…
- An important lesson from this selection is…
- In this selection, why did Francisco and Grandpa leave the weeds?
- This selection is not only about determination, it is also about…
- Why did the author have Grandpa and Francisco speak in Spanish?
What important advice would both Grandpa and Kate give?

In both reading selections you read about main characters who…

How are Francisco and Kate different?

How were the characters rewarded for showing determination and following through?
Applying Ideas to a Task

If you were trying to do something that was very hard, and you did not think you could get it done, would you keep trying or quit? Use examples from the two stories we read to support your decision.
Scoring

Answers questions by making connections between readings and using ideas from both readings to support position taken.

Answers questions and uses ideas from at least one story to support position taken.

Answers question and refers to ideas in one text.

Answers question or responds to theme.
The student clearly and effectively chooses key or important ideas from each reading selection to support a position on the question and to make a clear connection between the reading selections. The point of view and connection are thoroughly developed with appropriate examples and details. There are no misconceptions about the reading selections. There are strong relationships among ideas. Mastery of language use and writing conventions contributes to the effect of the response.
Bottom Line

Mixed model assessment along the lines of NAEP

- Some multiple choice
- Some short answer
- Some constructed response (real performance items)
- Some within text
- Some cross text
- Some big ideas
- Some details
- Lots of relationships among ideas
Why this model?

- Acknowledges the conceptual and psychometric contributions of different formats and the theories of comprehension that lie behind them.
- Admits that we have, at least at present, no conclusive evidence to direct us to the one best model of comprehension assessment.
- Maps onto some useful instructional activities.
The useful instructional activities that the mixed model maps onto

📚 Building a rich text base (what does it say?)
  ✓ Facts, relationships, inferences

📚 Building a model of what the text means (text filtered through prior knowledge)
  ✓ Reminders, comparisons, unstated details and motives

📚 Some analysis and critique
  ✓ What is the author up to? How is (s)he trying to shape my thinking?
And that combination seems…

Pretty consistent with a long line of research and theory development over the past century.

Now on to Vocabulary…
There is only one book in Books in Print with the title, Vocabulary Assessment

What field do you think it is in?

✓ Reading
✓ Oral Language Development
✓ Intelligence Testing
✓ English as a Foreign/Second Language
Vocabulary Portion of the talk

- Define domain of interest
- A short history of vocabulary assessment
- Some important features of the domain of vocabulary assessment
**Domain of interest**

**For sure:** Knowledge of word meanings and the conceptual networks in which they exist. BUT

✓ Which words?
  - in general
  - a set of specific words from a story, unit, book, etc.

**Perhaps:** The ability to use available cues, both inside words and outside of them, to infer--and maybe to learn--the meanings of words
A very short history of vocabulary assessment

Vocabulary assessment has been around as long as we have had:
- Assessment of any sort
- Reading assessment

Has always been a part of intelligence testing.

Has always been a part of reading assessment, on its own, and part of comprehension.

Has always been a major part of second language assessment.
Trends over time

Early on: test isolated words and find their synonyms or meanings

Not surprisingly, there has been a movement toward contextualization over time

- Psycholinguistic and cognitive revolution
- Constructivist pedagogies
  - Whole language and its kissing cousins
  - Communicative competence (ESL)
Circa 1920s thru 1950s

A _______ is used to eat with
1. Plow
2. Fork
3. Hammer
4. Needle

given a feature and asked to find a word that possesses it

Foolish
1. Clever
2. Mild
3. Silly
4. Frank

given a word and asked to find a rough paraphrase or synonym
He discovered a new route through the mountains.

1. Wanted
2. Found
3. Traveled
4. captured

Their success came about as a result of your assistance.

1. according to
2. before
3. because of
4. during

Note that context *does allow* us to assess *abstract* words.
1950s: deliberately
   1. Both
   2. Noticeably
   3. Intentionally
   4. Absolutely

1970s: He was found guilty because he did the act deliberately.

deliberately
   1. Both
   2. Noticeably
   3. Intentionally
   4. Absolutely
In a (1) democratic society, individuals are presumed innocent until proven guilty. The (2) establishment of guilt is often a difficult task. One consideration is whether or not there remains a (3) reasonable doubt that the suspected persons committed the act in question. Another consideration is whether the acts were committed (4) deliberately.

(4)
1. Both
2. Noticeably
3. Intentionally
4. absolutely

Compared to other formats, this one showed the highest reliability, predictive validity, discrimination
1995: among comprehension questions, insert vocabulary…

…Two reasons are usually advanced to account for this tardy development; namely the mental difficulties…

The word tardy in line 2 is closest in meaning to

1. Historical
2. Basic
3. Unusual
4. Late

Note: Still an open question of whether you report vocabulary separately
Now, in the age of on-line assessment...

The Southwest has always been a dry country, where water is scarce, but the Hopi and Zuni were able to bring water from streams to their fields and gardens through irrigation ditches. Because it is so rare, yet so important, water played a major role in their religion.

Look at the word rare in the passage. Click on the word in the text that has the same meaning.
NAEP’s likely influence

- That NAEP is in the game will elevate the role of vocabulary
- NAEP’s standards (achievement levels) and format will also influence assessment
- Good development to attend to.
- Note: whether it is a separate scale depends on
  - Resources (for item development)
  - Psychometrics (will it scale separately from comprehension)
What will be tested on NAEP?

- Assess words characteristic of written language not oral language
- Label generally familiar and broadly understood concepts, even if the words themselves are not familiar (akin to Isabel’s discussion of Tier 2):
  - Stunning but not pretty
  - Prosperous but not rich
  - Demonstrate but not show
- Are required to built a sensible rendition of the text (and preferably linked to central ideas in the text).
- Are characteristic of grade level material (4, 8, 12)

These are Beck & McKeown’s Tier Two words.
What won’t be tested on NAEP

Words that are **narrowly defined** and not widely used (appears to be Tier Three, technical vocabulary) or just **arcane** (hamlet or rivulet)

Words that label the **main idea of the text** (e.g., emancipation in Emancipation Proclamation)

Words that are part of **most students’ speaking** vocabulary (Tier One words)

Words with meanings that are **readily derived from context** (appositives, parenthetical definitions)
NAEP Distractor features

- Can present a more common meaning for the word, which must be ignored in favor of the meaning in the text.
- Can present correct information from the text that is NOT the meaning of the word.
- May be an alternative interpretation of the context in which the word occurs.
- Visually or auditorily similar words.

Note: Useful to have a theory of distractor generation because it gives meaning to errors.
NAEP Achievement levels for vocabulary
Advanced readers will have outstanding vocabularies,
✓ With a sound knowledge of words and terms beyond their grade level.
In addition, they will have an excellent grasp of the multiple meanings of an extensive set of words and complex networks of associations to the words they know.
They will also have a strong base of words that identify complex and abstract ideas and concepts.
Finally their sophistication with words and word meanings will enable them to be highly flexible in extending the senses of words they know
✓ to appropriately fit different contexts.
Proficient readers will have sizeable meaning vocabularies,
  ✓ including knowledge of many word and terms above grade level.

They will also have greater depth of knowledge of words, beyond the most common meaning.

Proficient readers will be flexible with word meanings and able to extend the senses of words whose meanings they know
  ✓ in order to appropriately fit different contexts and understand passage meaning.
Readers at the basic level will generally have limited, concrete vocabularies that consist primarily of words at and below grade level. Knowledge of these words will be limited to the most familiar definition, making it difficult to identify the appropriate meaning of a word among the distractors.
NAEP’s implicit theory

- Size
- Depth
- Contextual flexibility and situatedness
After John Read


You can assess vocabulary with an eye toward these distinctions:

- Discrete--->Embedded (phenomenon)
- Selective--->Comprehensive (corpus)
- Context independent--->context dependent (format)
Discrete: vocabulary as an independent construct
  ✓ (e.g., lots of standardized tests report a separate comprehension score)

Embedded: vocabulary is assessed but feeds into a score for a larger construct
  ✓ (e.g., added to comprehension aggregate score)*
  ✓ (e.g., on a typical test, you get
    • Comprehension
    • Vocabulary
    • Total reading

*What NAEP has always done
Selective--->Comprehensive

What is the “population” to which we want to generalize?

Selective: a targeted set of vocabulary items
- Those in the selection at hand
- Those in the unit at hand
- Those in a certain band of frequency (e.g., 1000 most frequent)

Comprehensive: all the words in some domain or performance
- All the words in an essay (when we rate the sophistication of an essay)
- All the words in a speech
- All the words in a text
- All the words in a corpus, language

For selective, we tend to measure the population; for comprehensive we tend to draw samples.

A Dilemma: the vocabulary samples in standardized tests are not a sample of anything in particular--rely on norms (how other kids do, not any “concept” of a domain)
Context dependence

Must you use the context to ascertain the meaning?

Context independent
- Don’t have it OR
- Don’t need it when you DO have it

Context dependent: can’t get the right answer without reference to context
- Rare words
- Nonsense words
- Words with multiple meanings
- Missing words
Context absent

consumed
1. Ate or drank
2. Prepared
3. Bought
4. Enjoyed
Context present but not critical

The people **consumed** their dinner

1. Ate or drank
2. Prepared
3. bought
4. enjoyed
The people **consumed** their dinner.
1. Ate or drank
2. Used up
3. Spent wastefully
4. Destroyed

The air conditioner **consumed** a lot of energy.
1. Ate or drank
2. Used up
3. Spent wastefully
4. Destroyed

Note: the distractor set matters!
Questions to ask in building or evaluating vocabulary assessments

- What does it mean to know a word?
- What counts as a word?
- How do we choose words?
- How do we know whether selected words are known?
Format issues

- Pictures make life easier…
  ✓ Reduce the reading load
- And harder…
  ✓ Ambiguity
- Distractors matter
  ✓ They largely determine the cognitive focus of the task
  ✓ What is varied is what you must pay the most attention to..
What is varied is what is assessed

Look at the following examples to determine the focus
Prosperous
a. wealthy
b. sad
c. tall
d. happy

Prosperous
a. doing well
b. in trouble
c. not very happy
d. very lucky
A person who is prosperous could be said to be…
✓ doing well financially
✓ in trouble
✓ not very happy with life
✓ very lucky

A person who is doing well financially could be said to be
✓ anxious
✓ sick
✓ open
✓ prosperous

Note that the second item permits you to infer the meaning from context IF you know the meanings of the words in the sentence and the other choices.
With time things got better and many settlers became prosperous.

✓ wealthy
✓ sad
✓ tall
✓ happy

With time things got better and many settlers became ____________

✓ anxious
✓ sick
✓ open
✓ prosperous
With time things got better and many settlers became ______________.
✓ prospered
✓ prosper
✓ prosperous
✓ prosperous
Assessing general vocabulary growth

- Usually we resort to some normative assessment: How much growth did they make compared to other kids in the norming sample?
  - PPVT, vocabulary subtest of any reading assessment,

- Logically possible, but not very practical, to draw samples from a very large corpus
  - Words in the band of frequency from 1-1000, 3000-5000, etc.
  - Words in a large but specifiable domain (a semester long course or in a big textbook)
New Validity Initiatives

- We must move away from normative benchmarks
  ✓ How do you stack up against others?
- To absolute standards of mastery
  ✓ What percentage of the words in this domain or population do you show mastery over?
- Not hard in the days of computerized assessment
  ✓ Carefully defined populations or domains
  ✓ Computer adaptive assessment
Bottom line

- We have a long history of assessing vocabulary, but…
- Not much research to guide us in selecting the perfect approach
- We can get along with many of the tools we have, but…
- We need some significant work on the construct validation of vocabulary assessments
- The research community, along with the publishing community, needs to provide teachers with better tools.
- Don’t forget technology