Reading Comprehension Assessment: One VERY Resilient Phenomenon

P. David Pearson, UC Berkeley
Diane Hamm, MSU
Key references


Purpose

- Build an argument for a fresh line of inquiry into the assessment of reading comprehension
- By providing a rich and detailed historical account of reading comprehension, both as a theoretical phenomenon and an operational construct
- Bonus: Try to offer my explanation of why the short passage, multiple-choice exam format is so resilient
Why now?

- Renewed interest among scholars
  - Rand report
- Uneasiness among practitioners that the code, as important as it is, may not be the point of reading
- The most important outcome of reform
- National thirst for accountability requires impeccable measures (both conceptually and psychometric)
- Pleas of teachers desperate for useful tools (need a tool that does for RC what running records does for word id)
Reading comprehension assessment has always vexed researchers

- 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
  - Quiz them on the details
  - Request the big ideas
Most of the measures interpose some other skill or capacity between the act and the evidence

- Writing
- Talking
- The conventions of multiple-choice assessments
- These interposed processes inevitably compromise our capacity to draw inferences about comprehension both as a generic and a passage specific enterprise
History nurtures modesty

Just about any approach to assessing reading comprehension that has arisen in the last 30 years has a precedent that is at least 70 years old.
Caveats

I apologize, in advance, for all the studies, especially those done by people in this room, that I will fail to cite.

I also want to make it clear that I am limiting myself to READING COMPREHENSION assessment, not prerequisites to or correlates of RC.

Just because I chronicle a practice does not mean that I advocate it!
The Century old Roots of Reading Comprehension Assessment

- Fit the context of the period just after the last turn of the century
- Curricular shift from oral to silent reading
  - From Accuracy and Expressiveness (oratory and declamation) of oral reading
  - To Indicators of understanding
The Roots…

- Emerging Scientism in education
- Growing numbers of students
- Changing demographics: widespread illiteracy
- Efficient means of assessing (e.g. Army Alpha Exams)
- Move from high to a low inference index
1895 Binet

- Used RC tasks as a measure of IQ not reading
- Not inconsistent with Binet’s notion that IQ should index school-based reasoning capacity
Every one of us, whatever our speculative opinion, knows better than he practices, and recognizes a better law than he obeys.”

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 IQ test, cited in Johnston, 1984, (undated)
1916 Kansas Silent Reading Test*

- Kelley
- “fill in the blanks”
- some verbal logic problems
- some procedural tasks
- Complete as many of 16 tasks as possible in a limited time

*The first published standardized comprehension test.
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
Touton and Berry (1931) Error analyses

(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

- Starch: relevant words recalled as a function of total words recalled
- Courtis (1914): words remembered/words in text
- Chapman (1924): Find the words in part 2 that do not fit the words in part 1 of the paragraph.
- Note similarity to later free recall and error detection models of assessment
Enter Psychometrics in the late 1930s

- 1935: IBM introduced the IBM 805 scanner
- 1935: Kelley: Factor Analysis
- 1944: Davis: Fundamental Factors
### 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></td>
</tr>
<tr>
<td>Follow passage organization</td>
<td>Draw inferences about content</td>
</tr>
<tr>
<td>Main thought</td>
<td>Literary devices</td>
</tr>
<tr>
<td>Answer specific text-based questions</td>
<td>Author’s purpose</td>
</tr>
</tbody>
</table>

Word factor and a reasoning factor
Other Factor Analyses

- Harris 1948: found a single factor
- Derrik (1953) found 3
- 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)
### Davis 1972

<table>
<thead>
<tr>
<th>1. Remembering word meaning</th>
<th>5. Drawing inferences from the content</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Word meanings in context</td>
<td>6. Recognizing the author’s tone and mood and purpose</td>
</tr>
<tr>
<td>3. Understanding content stated explicitly</td>
<td>7. Recognizing literary techniques</td>
</tr>
<tr>
<td>4. Weaving together ideas in the content</td>
<td>8. Following the structure of the content</td>
</tr>
</tbody>
</table>
Davis 1972

- Remembering word meanings
- Drawing inferences from content
- Structure of the passage
- Writerly techniques
- Explicit comprehension
- Put an end to factor analytic studies
Cloze Procedure

- Wilson Taylor (1953): every 5th word
- Bormuth (1966): the basis of readability research
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 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
Still survives

- DRP
- Stanford Diagnostic
- ESL
Passage Dependency

\[ P_{\text{passage}} - P_{\text{isolation}} \]

A quiet stir in the late 60s and early 70s (Tuinman)
Died in the wake of Schema Theory’s embrace of prior knowledge
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
CRT takes over

- Wisconsin Design for Reading Skill Development
- Fountain Valley
- Virtually every basal program by the mid 1970s
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.

(adapted from the Ginn Reading Program, 1982)
Reactions to this movement

- Provided fuel for the constructivist reforms that were gathering momentum
- Died in the early 90s basals for about 6 years
- Only to be revived recently

Domain referenced assessment

- John Bormuth, Toward a Theory of Achievement Test Items
- Identify the domain as texts
- Map all of the logical relations among sentences.
- Using linguistic transformations, develop all possible Wh questions --> items
- Randomly sample from the domain
- Survives in Math, not reading
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
Contrasts between what we know and what we do

New views of the reading process tell us that . . .

Yet when we assess reading comprehension, we . . .

<table>
<thead>
<tr>
<th>Prior knowledge is an important determinant of reading comprehension.</th>
<th>Mask any relationship between prior knowledge and reading comprehension by using lots of short passages on lots of topics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A complete story or text has structural and topical integrity.</td>
<td>Use short texts that seldom approximate the structural and topical integrity of an authentic text.</td>
</tr>
<tr>
<td>The diversity in prior knowledge across individuals as well as the varied causal relations in human experiences invites many possible inferences to fit a text or question.</td>
<td>Use multiple-choice items with only one correct answer, even when many of the responses might, under certain conditions, be plausible.</td>
</tr>
<tr>
<td>Inference is an essential part of the process of comprehending units as small as sentences.</td>
<td>Rely on literal comprehension test items.</td>
</tr>
<tr>
<td>The ability to vary reading strategies to fit the text and the situation is one hallmark of an expert reader.</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td></td>
</tr>
<tr>
<td>Seldom assess how and when students vary the strategies they use during normal reading, studying, or when the going gets tough.</td>
<td></td>
</tr>
</tbody>
</table>
A sense that we had

- Paid too much attention to measurement theory and
- Not enough to reading theory
Structural representations

- Used in test development
- Determine hierarchical and sequential relations
- A theory of importance
- Determines which nodes should be assessed
Authentic Texts

- Select, not construct, texts for understanding
  - (started a cottage industry for magazine publishers)
- Can’t tinker with the text to rationalize items and distractors
  - (drove professional item writers crazy)
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.
Rate all of the responses on some scale of relevance

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.

Best predictor of retelling scores
Include

- Complex indicators of comprehension
- Prior knowledge
- Metacognition
- Habits, attitudes, and dispositions
Most notable examples

- MEAP (Wixson, Peters, Paris)
- IGAP (Valencia, Shanahan, Pearson, Reeve)
Some findings from IGAP

- When we plugged in Comprehension, Prior Knowledge, Metacognition, Habits/Attitude

- We emerged with these factors
  - metacognitive
  - habits/attitudes items
  - a combination of the comprehension and prior knowledge items
Fate

- Went the way of all tests that challenge the conventional wisdom
- Not good to teach to (e.g. metacognitive items)
- Went down in the mid 1990s when they tried to add on an individual score reporting component
Sentence verification task

- Original: Verbatim repetition of a sentence in the passage
- Paraphrase: The same meaning as an original but with lots of semantic substitutes for words in the original sentence.
- Meaning change: Uses some of the words in the passage but in a way that changes the meaning of the original sentence.
- Distractor: A sentence that differs in both meaning and wording from the original.
Judge each as old or new

- Most people seem content with polyester fillings and such. (Original)
- You don't know what comfort is until you've sunk your head into 3,000 bits of polyester. (Meaning change)
- It is always fun visiting grandparents because they take you someplace exciting, like the zoo or the circus. (Distractor)
- Being able to hear stories of when his mom and dad were kids was one of the great things about having grandparents around, Tim concluded. (Paraphrase)
- His favorite grandparent was his mother's mother. (Distractor)
Why isn’t SVT used more often?

- Does not pass the prima facie test
- Time consuming to prepare
Sociocultural and Literary Perspectives

- 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,
Another CLAS

Now you will be working in a group. You will be preparing yourself to do some writing later. Your group will be talking about the story you read earlier. A copy of the story is provided before the group questions if you need to refer to it. Some of the activities in this section may direct you to work alone and then share with your group, and other activities may have all of you working together. It is important to take notes of your discussion because you will be able to use these notes when you do your writing.

Read the directions and do the activities described. Members of the group should take turns reading the directions. The group leader should keep the activities moving along so that you finish all activities.

You’ll have 15 minutes for these prewriting activities.
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

The legacy:
- Mixed models
- Classroom assessment
Sharing thinking

- Think alouds
  - Olshavsky
  - Hartman
  - NVT

- Write alongs
  - Farr & Greene
  - CLAS
NAEP

1960s: Goal free evaluation

What you see is what you get
NAEP 1970s

- 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.
NAEP 1980s

- value reading and literature
- comprehend written works
- respond to written works in interpretive and evaluative ways
- apply study skills
NAEP 1990s

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 _____

DEMONSTRATE CRITICAL STANCE
- What could be added to improve the author’s argument
NAEP concerns

- The framework does not pass psychometric muster
- Not much information at the lower end of the performance scale (no floor)
- Item format: Do CR items add any value to the information gained?

Not if they are MC in disguise?
Mapping test scores and text difficulty on the same scale

- Stenner et al: Lexile
- DRP
- Carver
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?
What does it mean to achieve a given comprehension score?

- Find a population of kids with a narrow band of overall comprehension scores
- Administer lots of subskill tests, decoding, vocabulary, and comprehension
- Evaluate prerequisiteness and compensatory hypotheses
  - Which types of knowledge/skill are essential
  - How many ways are there to get to 6.5?
- Valencia study (later this morning)
- Note that New Standards Reference Exam privileges compensatory concept.
More questions to answer?

For accommodations, how do we weigh increased participation against potential sources of invalidity?

- Time
- Glossary
- Mode of presentation
Starting over

Go back to a set of theoretical conceptualizations of comprehension
- Component Skills
- Knowledge Driven models (Schema Theory and Construction-Integration)
- Contextually Driven models (Socio-cultural or critical)
- Executive Control models (metacognition and Cognitive Flexibility Theory)

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
More steps

- Develop a “gold standard” for comprehension—how do we get as close as possible to that ineffable phenomenon?
- My candidate: Some on-line assessment of both the content (ideas in text) and the affect (phenomenological sense) of comprehension (akin to the write alongs)
- Examine the predictive validity of the assessment models generated from each theoretical perspective in relation to the gold standard
- Be open to the possibility of a mixed model
Conclusion

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

- 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.
- To serve the needs of both classroom teachers and policy makers.
- A welcome challenge.